Capitolo 5
Stima parametrica

per gli studenti del corso di
Stima e identificazione

Luigi Chisci, 28 Marzo 2020

Generalità sui problemi di stima

Per stima, genericamente parlando, si intende il processo di inferire il valore di una variabile X di interesse dall'osservazione di un'altra variabile Y che dipenda in qualche modo da X. Pertanto i problemi di stima coinvolgono due attori:

- la variabile da stimare $X \in \mathbb{R}^n$;
- la variabile osservata $Y \in \mathbb{R}^p$.

Si fa notare come la variabile osservata Y sia di norma il risultato di un esperimento casuale di misura e pertanto considerata come variabile aleatoria. Viceversa, la variabile da stimare X, a seconda dell'approccio adottato, può essere interpretata come grandezza deterministica di valore ignoto oppure anch'essa come variabile aleatoria (approccio Bayesiano).

Si definisce *stimatore di $X \in \mathbb{R}^n$ basato sull'osservazione $Y \in \mathbb{R}^p$* una generica funzione $g : \mathbb{R}^p \to \mathbb{R}^n$ operante sull'osservazione Y per produrre una stima $\widehat{X} = g(Y)$ di X. È opportuno distinguere fra la *variabile aleatoria stima* $\widehat{X} = g(Y)$ e la *stima puntuale* $\hat{x} = g(y)$ risultante dall'applicazione dello stimatore $g(\cdot)$ al particolare valore y della variabile aleatoria Y osservato durante l'esperimento di misura effettuato. Si definisce anche l'*errore di stima* $\widetilde{X} = X - \widehat{X} = X - g(Y)$ come la differenza fra la variabile da stimare e la sua stima. Un obiettivo naturale è quello di rendere tale errore di stima "piccolo" in accordo a qualche criterio, deterministico o probabilistico, da precisare opportunamente. A tale proposito, un criterio valido potrebbe essere quello di avere un errore di stima a media nulla (*stimatore non polarizzato*), i.e.,

$$E\left[\widetilde{X}\right] = E\left[X - \widehat{X}\right] = 0 \implies E\left[\widehat{X}\right] = E\left[X\right]$$

e di varianza $E\left[\widetilde{X}\widetilde{X}^T\right]$ minima.

Un ingrediente fondamentale dei problemi di stima è il *modello di osservazione* (o di misura) che esprime matematicamente la dipendenza della variabile osservata Y dalla variabile di interesse X. Il modello di osservazione è in generale descritto da un'*equazione di misura* della forma

$$Y = h(X, V) \tag{1.1}$$

in cui compare un'ulteriore variabile $V \in \mathbb{R}^q$, detta *errore di misura*, a tenere conto dell'inevitabile errore che si commette nella misura di una qualunque grandezza reale. Di norma, V è considerata come variabile aleatoria con PDF $f_V(\cdot)$. Si assume che V abbia media nulla e varianza $R = R^T > 0$; inoltre, nel caso Bayesiano, in cui la variabile da stimare X è considerata aleatoria, si suppone che V sia incorrelato con X.

Si fa notare come l'ipotesi di media nulla non limita in alcun modo la generalità: se infatti l'errore di misura $\mathcal{V}$ avesse media $\overline{v} \neq 0$, si potrebbe ridefinire $V = \widetilde{\mathcal{V}} \stackrel{\triangle}{=} \mathcal{V} - \overline{v}$ come errore di misura a media nulla ed inglobare la media $\overline{v}$ nella funzione di misura $h(\cdot, \cdot)$ se la media è nota, oppure nel vettore da stimare X se tale media è incognita. In termini più precisi, dato il modello di misura $Y = H(\mathcal{X}, \mathcal{V}]$ con errore di misura $\mathcal{V} \sim (\overline{v}, R)$ ci si può sempre ricondurre al modello di misura (1.1) con errore di misura $V \sim (0, Q)$ nel seguente modo:

- se $\overline{v} = 0$, allora
$$X = \mathcal{X}, \quad V = \mathcal{V}, \quad h(X, V) = H(X, V);$$
- se $\overline{v} \neq 0$ è nota, allora
$$X = \mathcal{X}, \quad V = \mathcal{V} - \overline{v}, \quad h(X, V) = H(X, V + \overline{v});$$
- se $\overline{v} \neq 0$ non è nota, allora
$$X = \left[\mathcal{X}^T, \overline{v}^T\right]^T, \quad V = \mathcal{V} - \overline{v}, \quad h(X, V) = H\left([I_n, 0]X, V + [0, I_q]X\right).$$

In molte situazioni pratiche, è comune considerare un errore di misura additivo per cui (1.1) si riduce al seguente *modello di osservazione additivo*:

$$Y = h(X) + V \tag{1.2}$$

con *funzione di misura* $h(\cdot)$ dipendente dalla sola variabile X ed errore di misura $V \in \mathbb{R}^p$ della stessa dimensione della variabile osservata Y. In situazioni ancora più particolari, comunque ricorrenti nella pratica ingegneristica, la funzione di misura risulta lineare, i.e. $h(X) = CX$ per una opportuna matrice $C \in \mathbb{R}^{p \times n}$, e (1.2) si riduce pertanto al seguente *modello di osservazione lineare*:

$$Y = CX + V. \tag{1.3}$$

A conclusione di questo paragrafo, si vogliono illustrare alcuni esempi pratici di problemi di stima al duplice scopo di: (1) chiarire il significato dei vari ingredienti di un problema

di stima; (2) mostrare come tali problemi abbiano applicazione nei più svariati contesti dell'ingegneria e della scienza.

Esempio 1: Stima della legge oraria - Come è ben noto, per *legge oraria* si intende la relazione matematica che lega una certa grandezza fisica $x(t)$ (e.g., la posizione di un oggetto) al tempo t. È prassi comune considerare una legge oraria polinomiale di opportuno grado $n-1 \geq 0$, i.e.

$$x(t) \;=\; X_1 + X_2\, t + \cdots + X_n\, t^{n-1} \;=\; \sum_{i=1}^{n} X_i\, t^{i-1} \;=\; \left[1, t, \ldots, t^{n-1}\right]\; X \tag{1.4}$$

dove $X \stackrel{\triangle}{=} [X_1, \ldots, X_n]^T \in \mathbb{R}^n$ è il vettore dei parametri incognito della legge oraria. A fini previsionali, è di fondamentale importanza stimare il vettore dei parametri X della legge oraria. A tale proposito si possono effettuare osservazioni sperimentali, $y(t)$, della variabile $x(t)$ ad un certo numero p di istanti di osservazione $t_1 < t_2 < \cdots < t_p$. Poichè l'osservazione differirà dalla variabile di interesse per effetto di errori di misura, si avrà una relazione del tipo

$$y(t) \;=\; x(t) + v(t) \;=\; \left[1, t, \ldots, t^{n-1}\right]\; X + v(t).$$

Raccogliendo tutte le osservazioni in un vettore $Y = [Y_1, \ldots, Y_p]^T = [y(t_1), \ldots, y(t_p)]^T \in \mathbb{R}^p$ si ottiene un modello di osservazione lineare della forma (1.3) con

$$C = \begin{bmatrix} 1 & t_1 & \cdots & t_1^{n-1} \\ 1 & t_2 & \cdots & t_2^{n-1} \\ \vdots & & & \vdots \\ 1 & t_p & \cdots & t_p^{n-1} \end{bmatrix}, \quad V = \begin{bmatrix} v(t_1) \\ v(t_2) \\ \vdots \\ v(t_p) \end{bmatrix} \sim \left(0, R = \sigma_v^2 I_n\right)$$

nell'ipotesi ragionevole in cui tutti gli errori di misura $v(t_1), \ldots, v(t_p)$ siano fra loro indipendenti, a media nulla ed abbiano la stessa varianza σ_v^2. Si noti come la stima $\widehat{X}$ consentirebbe la previsione della grandezza di interesse ad ogni istante t successivo all'ultimo istante di osservazione, mediante la relazione

$$\hat{x}(t|t_p) \;=\; [1, t, \ldots, t^{n-1}]\widehat{X}, \qquad \forall t > t_p.$$

A questo proposito, sarebbe di notevole interesse poter procedere ricorsivamente in modo da disporre, ad ogni istante temporale t, di una stima $\widehat{X}_i$ di X basata su tutte le osservazioni precedenti all'istante stesso, i.e. $t_1 < t_2 < \cdots < t_i < t$, e aggiornare tale stima al valore $\widehat{X}_{i+1}$ all'istante t_{i+1} quando si rende disponibile l'osservazione successiva $Y_{i+1} = y\,(t_{i+1})$. Si noti che, in questo caso, $\widehat{X}_i$ rappresenta la stima di X basata su

$Y_{1:i} \triangleq [Y_1, \ldots, Y_i]^T = [y(t_1), \ldots, y(t_i)]^T$ e che l'approccio ricorsivo sopra delineato permetterebbe di affrontare in modo efficace situazioni in cui nuove osservazioni arrivano sequenzialmente e continuativamente allo stimatore.

Naturalmente è possibile generalizzare la legge oraria in modo da considerare arbitrarie funzioni del tempo $f_1(t), f_2(t), \ldots, f_n(t)$ al posto delle specifiche funzioni $f_1(t) = 1, f_2(t) = t, \ldots, f_n(t) = t^{n-1}$ considerate in (1.4). Questo non altera in alcun modo la natura lineare del problema di stima che rimane lo stesso salvo la ridefinizione della matrice di misura C in (1.3) come

$$C = \begin{bmatrix} f_1(t_1) & f_2(t_1) & \cdots & f_n(t_1) \\ f_1(t_2) & f_2(t_2) & \cdots & f_n(t_2) \\ \vdots & & & \vdots \\ f_1(t_p) & f_2(t_p) & \cdots & f_n(t_p) \end{bmatrix}.$$

Le applicazioni della stima della legge oraria sono molteplici, includendo ad esempio: la previsione del moto di un oggetto balistico, l'interpolazione di una funzione con punti acquisiti sperimentalmente, la stima di ampiezza e fase di un segnale sinusoidale di frequenza nota, etc.. □

Esempio 2: Localizzazione GPS - Un sistema di navigazione satellitare - quale ad esempio GPS (*Global Positioning System*) gestito dal governo degli Stati Uniti d'America ma anche la sua controparte russa GLONASS o quella europea GALILEO - consiste di una rete dedicata di satelliti artificiali mediante la quale si fornisce ad un terminale mobile o ricevitore GPS informazioni sulle sue coordinate geografiche e sull'orario. Più precisamente il ricevitore GPS riceve da ogni satellite visibile i un messaggio che specifica la posizione, nelle tre coordinate Cartesiane latitudine-longitudine-altitudine (ξ_i, η_i, ζ_i), del satellite nonché l'istante temporale $t_{i,tx}$, secondo un orologio atomico ultra-preciso presente sul satellite, al quale tale messaggio è stato trasmesso. A sua volta, il ricevitore GPS rileva l'istante di ricezione $t_{i,rx}$ di tale messaggio, secondo un proprio orologio assai meno preciso e non sincronizzato con quello del sistema satellitare, e conseguentemente calcola una distanza satellite-ricevitore Y_i mediante la relazione $Y_i = c\,(t_{i,rx} - t_{i,tx})$ dove c è la velocità di propagazione nota del segnale radio circa pari alla velocità di propagazione della luce nel vuoto, i.e. $c \cong 3 \cdot 10^8\ [m/s]$. In realtà, la singola osservazione GPS Y_i fornisce una misura molto grossolana della distanza fra satellite i-esimo della costellazione GPS e ricevitore GPS nel senso che tale misura è affetta da diverse sorgenti di errore, in primis l'errore di sincronizzazione degli orologi Δt che induce un corrispondente errore di distanza $\Delta r = c\,\Delta t$ di notevole entità (e.g. $\Delta t = 1\ [ms]$ implica $\Delta r = 300\ [km]$). Per questo motivo, Y_i è detta *pseudo-distanza*. Indicate con (ξ, η, ζ) le coordinate Cartesiane di posizione (latitudine, longitudine ed altitudine) del navigatore GPS da localizzare e con p il numero di satelliti visibili dal navigatore GPS, le osservazioni (pseudo-distanze)

sono espresse dalle seguenti relazioni

$$Y_i \;=\; \sqrt{(\xi-\xi_i)^2+(\eta-\eta_i)^2+(\zeta-\zeta_i)^2} \;+\; c\,\Delta t \;+\; V_i \qquad i=1,2,\ldots,p$$

dove V_i tiene conto di altre sorgenti di errore di minore entità (rispetto all'errore derivante dalla desincronizzazione degli orologi), quali ad esempio effetti della relatività nonché della rifrazione del segnale radio in ionosfera e troposfera, etc.. Di norma, ma non necessariamente, gli errori V_i sono ipotizzati a media nulla e fra loro incorrelati. Risulta chiaro da quanto sopra esposto che la localizzazione GPS necessariamente coinvolge, oltre al vettore di posizione $[\xi,\eta,\zeta]^T$ del navigatore GPS, un'ulteriore quarta variabile Δt, errore di sincronizzazione degli orologi. Pertanto, in questo problema, la variabile da stimare è $X=[\xi,\eta,\zeta,\Delta t]^T \in \mathbb{R}^4$, di dimensione $n=4$, mentre la variabile osservata $Y=[Y_1,Y_2,\ldots,Y_p]\in\mathbb{R}^p$, di dimensione p pari al numero di satelliti in vista, è legata ad X dal seguente modello di osservazione:

$$\left\{\begin{array}{rcl} Y &=& h(X)+V \\ h(X) &=& [h_1(X),h_2(X),\ldots,h_p(X)]^T \\ h_i(X) &=& h_i\left(\xi,\eta,\zeta,\Delta t\right)=\sqrt{(\xi-\xi_i)^2+(\eta-\eta_i)^2+(\zeta-\zeta_i)^2}\;+\;c\,\Delta t \\ V &=& [V_1,V_2,\ldots,V_p]^T\sim\left(0,diag\left\{\sigma_1^2,\sigma_2^2,\ldots,\sigma_p^2\right\}\right) \end{array}\right. \tag{1.5}$$

Si fa notare come la costellazione di satelliti GPS sia configurata in modo tale che in ogni punto della Terra risultano visibili $p\geq 4$ satelliti così che siano sempre disponibili almeno quattro osservazioni per stimare le quattro variabili di interesse $\xi,\eta,\zeta,\Delta t$ oppure per quest'ultima, in alternativa a Δt, il corrispondente errore sulla distanza $\Delta r \stackrel{\triangle}{=} c\,\Delta t$. □

Esempio 3: Stima dello stato di una rete di distribuzione dell'energia elettrica - La stima dello stato di una rete di distribuzione dell'energia elettrica costituisce un problema di sempre maggiore importanza nella gestione delle reti moderne (*smart grid*), anche in relazione alle emergenti esigenze del mercato dell'energia nonché per scopi di sicurezza. Lo stato da stimare è costituito dai moduli delle tensioni nei vari *punti di connessione (bus=nodi)* della rete e dagli sfasamenti di tali tensioni rispetto ad un *bus di riferimento*. Precisamente, se la rete ha N bus, la variabile da stimare è

$$X=[V_1,V_2,\ldots,V_N,\theta_{2,1},\theta_{3,1},\ldots,\theta_{N,1}]^T\in\mathbb{R}^{2N-1},$$

di dimensione $n=2N-1$, dove V_i è il modulo della tensione al bus i-esimo mentre $\theta_{i,1}$ $(i=2,\ldots,N)$ rappresenta l'angolo di sfasamento del bus i rispetto al bus 1 di riferimento. Le osservazioni disponibili includono misure affette da errore delle seguenti grandezze elettriche:

- potenze attiva P_i e reattiva Q_i iniettate nel bus i;

- flussi di potenza attiva P_{ij} e reattiva Q_{ij} nelle varie linee di trasmissione (archi della rete) fra bus i e j;
- modulo della tensione V_i al bus i.

La variabile osservata $Y \in \mathbb{R}^p$ è di norma costituita da un sottoinsieme di cardinalità $p > n = 2N - 1$ delle suddette grandezze misurate con opportuni strumenti (*voltmetri, wattmetri*) in un certo numero (non necessariamente tutti) di punti di connessione (nodi) e linee di trasmissione (archi) della rete. Il modello di osservazione che lega X a Y utilizza le equazioni di *load flow* comunemente impiegate nell'ambito dell'ingegneria elettrica per il calcolo dei flussi di potenza in una generica rete elettrica. Tali equazioni risultano non lineari e nella forma di un modello additivo $Y = h(X) + V$ con funzione di misura $h(\cdot)$ dipendente dalla topologia della rete e da opportuni parametri (resistenze e reattanze di linea dei vari collegamenti nonché conduttanze e suscettanze dei vari nodi). Di norma, si ipotizza che l'errore di misura V abbia media nulla e matrice di covarianza diagonale, i.e. $V \sim \left(0, diag\left\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2\right\}\right)$ per valori opportuni delle varianze $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$. □

Esempio 4: Calibrazione dei parametri di un sistema dinamico - In molte circostanze si avverte l'esigenza di calibrare i parametri del modello matematico di un certo sistema reale di interesse ad osservazioni sperimentali sul sistema stesso. Tale modello si presenta normalmente sotto forma di sistema dinamico a tempo-continuo

$$\left\{ \begin{array}{rcl} \dot{s}(t) & = & f\left(t, s(t), \theta\right) \\ y(t) & = & h\left(s(t)\right) + v(t) \end{array} \right. \tag{1.6}$$

con vettore dei parametri da calibrare (stimare) $\theta \in \mathbb{R}^n$. Si suppone che, per effettuare la calibrazione (stima) dei parametri, siano disponibili osservazioni sperimentali rumorose $Y_i \stackrel{\triangle}{=} y(t_i) = h\left(s(t_i)\right) + v(t_i)$ agli istanti $t_1 < t_2 < \cdots < t_p$. In questo problema di stima, la variabile da stimare è $X = \theta$ mentre la variabile osservata è $Y = [Y_1, Y_2, \ldots, Y_p]^T = [y(t_1), y(t_2), , \ldots, y(t_p)]^T \in \mathbb{R}^p$. Il modello di osservazione è additivo della forma (1.2) con

$$\begin{array}{rcl} h(X) & = & [h_1(X), h_2(X), \ldots, h_p(X)]^T \\ h_i(X) & = & h\left(s(t_i)\right) \\ V & = & [v(t_1), v(t_2), \ldots, v(t_p)]^T \sim (0, \sigma_v^2 I_p) \end{array}$$

dove $s(t_i)$, dipendente da X, è la soluzione all'istante t_i dell'equazione differenziale parametrica $\dot{s}(t) = f(t, s(t), X)$ con opportuna condizione iniziale $s(t_0)$, $t_0 < t_1$. Nel caso in cui la condizione iniziale $s(t_0)$ dell'esperimento non sia nota, si può sostituire la soluzione $s(t_i)$ dell'equazione differenziale con la predizione $\hat{s}\left(t_i | Y_{1:i-1}\right)$ di $s(t_i)$ basata sulle osservazioni precedenti $Y_{1:i-1} \stackrel{\triangle}{=} \{y(t_1), \ldots, y(t_{i-1})\}$. La determinazione di tale predizione è un

problema di stima dello stato di un sistema dinamico che verrà studiato in dettaglio nel prossimo capitolo.

A titolo di esempio si considera la calibrazione dei parametri di un modello ambientale di crescita logistica di una specie. Il modello si presenta in questa forma

$$\begin{cases} \dot{s}(t) &= r\, s(t) \left[1 - \dfrac{s(t)}{K}\right] \\ y(t) &= s(t) + v(t) \end{cases} \tag{1.7}$$

con stato scalare $s(t)$ che rappresenta la densità di popolazione della specie (in $[kJ/m^2]$) e parametri da calibrare $X = [r, K]^T$, dove $r > 0$ è il tasso di crescita e $K > 0$ la capacità portante dell'ecosistema. □

Esempio 5: Ricostruzione di un'immagine affetta da rumore "speckle" moltiplicativo - Gli esempi precedenti facevano tutti riferimento a modelli di osservazione additivi (in particolare, l'esempio 1 ad un modello lineare). Di seguito, si vuole illustrare una situazione in cui il problema di stima coinvolge un modello *non additivo*, per la precisione *moltiplicativo*. Il problema in oggetto riguarda la ricostruzione di un'immagine affetta da rumore di tipo *speckle*. Quest'ultimo, detto in italiano rumore *sale e pepe*, si manifesta con un insieme di *macchioline* disposte casualmente sull'immagine per effetto del passaggio di un'onda coerente attraverso un mezzo disordinato. Si indichi con $X = [X_1, X_2, \ldots, X_n]^T \in \mathbb{R}^n$ l'immagine da ricostruire, rappresentata da un vettore in cui i pixel X_i per $i = 1, 2, \ldots, n$ sono disposti in ordine lessicografico. Se si assume che non ci sia perdita di pixel, l'immagine osservata ha la stessa dimensione, i.e. $Y = [Y_1, Y_2, \ldots, Y_n]^T$. Ciascun elemento di Y è il prodotto del corrispondente elemento di X per il corrispondente elemento dell'errore di misura V, i.e.,

$$Y_i \;=\; X_i\, V_i$$

ovvero, in forma vettoriale, mediante la moltiplicazione componente-a-componente

$$Y \;=\; h(X, V) \;=\; X \cdot V \tag{1.8}$$

fra l'immagine originaria X ed il vettore $V = [V_1, V_2, \ldots, V_n] \in \mathbb{R}^n$. Tipicamente, si ipotizza che il rumore speckle moltiplicativo V abbia componenti indipendenti e sia distribuito alla Rayleigh, i.e.,

$$\begin{aligned} V \sim f_V(v) &= \prod_{i=1}^{n} f_{V_i}(v_i) \\ F_{V_i}(v_i) &= \frac{v_i}{\sigma^2}\, \exp\left(-\frac{v_i^2}{2\sigma^2}\right) \qquad \sigma^2 > 0. \end{aligned}$$

Si noti che la distribuzione di Rayleigh ha media non nulla $\overline{v} = \sigma\sqrt{\pi/2}$ e varianza $\sigma_v^2 = (2 - \frac{\pi}{2})\sigma^2$. □

Approccio Bayesiano

Secondo l'approccio Bayesiano si considera X come variabile aleatoria ed il problema della stima di X dall'osservazione Y come determinazione della PDF condizionata di X dato Y, i.e. $f_{X|Y}(x|y)$. Si noti come la conoscenza di tale PDF condizionata fornisca informazione preziosa ed esauriente sulla variabile da stimare X. Ad esempio, fissato un qualunque sottoinsieme S di $\mathbb{R}^n$ è possibile valutare la probabilità che la variabile di interesse X appartenga ad S mediante l'integrazione

$$Prob\,(X \in S) \;=\; \int_S f_{X|Y}(x|y)\,dx.$$

Inoltre, è possibile estrarre da $f_{X|Y}(\cdot|\cdot)$ una stima di X in accordo a diversi possibili criteri, quali ad esempio:

- il criterio della *media condizionata* (stima/stimatore EAP= *Expected A Posteriori*) che dà luogo alla variabile aleatoria stima

$$\widehat{X}_{EAP} \;=\; E\,[X|Y] \tag{1.9}$$

  e corrispondentemente alla stima puntuale

$$\hat{x}_{EAP}(y) \;=\; E\,[X|Y=y] \;=\; \int x\, f_{X|Y}(x|y)\,dx; \tag{1.10}$$

- il criterio della *moda condizionata* o della *massima probabilità a-posteriori* (MAP= Maximum A-posteriori Probability) che fornisce la stima

$$\hat{x}_{MAP}(y) \;=\; \arg\max_{x\in\mathbb{R}^n} f_{X|Y}(x|y) \;=\; \arg\max_{x\in\mathbb{R}^n} f_{Y|X}(y|x)\; f_X(x). \tag{1.11}$$

Si fa notare come nel caso Gaussiano, i.e. quando la PDF condizionata $f_{X|Y}(\cdot|\cdot)$ risulta Gaussiana, media (baricentro) e moda (punto di massimo) di tale PDF coincidono e, di conseguenza, si ha $\hat{x}_{EAP}(y) = \hat{x}_{MAP}(y)$. Questa situazione si verifica, ad esempio, nel caso di modello di osservazione lineare $Y = CX + V$ con X, V Gaussiani ed incorrelati, i.e.

$$\begin{bmatrix} X \\ V \end{bmatrix} \;=\; \mathcal{N}\left(\begin{bmatrix} \overline{x} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_X & 0 \\ 0 & \Sigma_V \end{bmatrix}\right).$$

Infatti, in questo caso X e Y risultano congiuntamente Gaussiane, precisamente

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} I & 0 \\ C & I \end{bmatrix} \begin{bmatrix} X \\ V \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \overline{x} \\ C\overline{x} \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_X C^T \\ C\Sigma_X & C\Sigma_X C^T + \Sigma_V \end{bmatrix} \right)$$

da cui, per l'invarianza della distribuzione Gaussiana rispetto al condizionamento, si ha

$$f_{X|Y}(x|y) = \mathcal{N}\left(x; \hat{x}(y), \widehat{\Sigma}_X\right)$$

con

$$\begin{aligned} \hat{x}(y) = \hat{x}_{EAP}(y) = \hat{x}_{MAP}(y) &= \overline{x} + \Sigma_{XY}\Sigma_Y^{-1}(y - \overline{y}) \\ &= \overline{x} + \Sigma_X C^T \left(\Sigma_V + C\Sigma_X C^T\right)^{-1}(y - C\overline{x}) \\ \widehat{\Sigma}_X &= \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T \\ &= \Sigma_X - \Sigma_X C^T \left(\Sigma_V + C\Sigma_X C^T\right)^{-1} C\Sigma_X \end{aligned}$$

Viceversa, se il modello di osservazione è non lineare e/o non Gaussiano, i due criteri di stima, EAP e MAP, forniscono in generale stime diverse.

Stimatori non Bayesiani

L'approccio Bayesiano che considera X come variabile aleatoria con assegnata, eventualmente poco informativa, distribuzione di probabilità a priori, non è l'unico possibile.

Stimatore di massima verosimiglianza (ML)

Un classico approccio non Bayesiano, che non presume alcuna PDF a priori $f_X(\cdot)$ di X e di conseguenza non può definirne alcuna PDF a posteriori $f_{X|Y}(\cdot|y)$, è quello della *massima verosimiglianza* (ML = Maximum Likelihood in inglese). Come dice il nome stesso, il criterio ML consiste nel massimizzare la *funzione di verosimiglianza* $f_{Y|X}(y|x)$ definendo la stima

$$\hat{x}_{ML}(y) = \arg\max_{x \in \mathbb{R}^n} f_{Y|X}(y|x) \tag{1.12}$$

come il valore $x \in \mathbb{R}^n$ di X più verosimile con l'osservazione $y \in \mathbb{R}^p$. Il confronto di (1.12) con (1.11) mette in evidenza il legame fra lo stimatore ML e lo stimatore MAP che, viceversa, massimizza il prodotto della funzione di verosimiglianza $f_{Y|X}(y|\cdot)$ per la PDF a priori $f_X(\cdot)$. In particolare, se la PDF a priori fosse costante su tutto $\mathbb{R}^n$, cioè *non informativa*, i due stimatori fornirebbero lo stesso risultato, i.e. $\hat{x}_{ML}(y) = \hat{x}_{MAP}(y)$. Naturalmente, non è possibile avere

$$f_X(x) = \varepsilon, \quad \forall x \in \mathbb{R}^n$$

dovendo essere $\int f_X(x)dx = 1$. Ciononostante è possibile definire, in molti modi diversi, una PDF non informativa $p_{ni}(\cdot)$ come limite, per un certo parametro scalare $r > 0$ che tende all'infinito, di una famiglia parametrica di PDF che, all'aumentare del parametro r, tendono ad assumere valori sempre più piccoli in una regione sempre più grande. Ad esempio, definita l'ipersfera $\mathcal{B}_r \stackrel{\triangle}{=} \{x : \|x\| \leq r\} \subset \mathbb{R}^n$ di raggio r ed il relativo volume $vol\,(\mathcal{B}_r) \stackrel{\triangle}{=} \int_{\mathcal{B}_r} dx$, si può considerare la famiglia di PDF uniformi

$$p_r(x) \stackrel{\triangle}{=} \begin{cases} \dfrac{1}{vol(\mathcal{B}_r)}, & x \in \mathcal{B}_r \\ 0, & x \notin \mathcal{B}_r \end{cases}$$

e definire la *PDF non informativa* come

$$p_{ni}(x) = \lim_{r\to\infty} p_r(x).$$

In modo alternativo, considerando una famiglia di PDF Gaussiane, si potrebbe adottare la definizione

$$p_{ni}(x) = \lim_{r\to\infty} \mathcal{N}(x; 0, rI_n).$$

Si fa notare come $p_{ni}(x)$ sia in qualche modo antitetica alla delta di Dirac $\delta(x)$, nel senso che quest'ultima assume valori infinitamente grandi in un sottoinsieme infinitamente piccolo di $\mathbb{R}^n$ mentre, al contrario, $p_{ni}(x)$ assume valori infinitamente piccoli in un sottoinsieme infinitamente grande di $\mathbb{R}^n$. Riassumendo, nel caso $f_X(x) = p_{ni}(x)$ che riflette la situazione in cui non si ha informazione a priori sulla variabile X da stimare, si ha coincidenza degli stimatori MAP e ML, vale a dire

$$f_X(x) = p_{ni}(x) \implies \hat{x}_{ML}(y) = \hat{x}_{MAP}(y).$$

In altri termini, lo stimatore non Bayesiano ML può essere re-interpretato come stimatore Bayesiano MAP con PDF a priori non informativa.

Stimatore ai minimi quadrati (LS)

Un ulteriore approccio classico non Bayesiano è quello ai *Minimi Quadrati* (in inglese: LS = Least Squares) che, per un modello di osservazione additivo $Y = h(X) + V$, si pone l'obiettivo di minimizzare una opportuna norma pesata dell'errore di misura $v = y - h(x)$, precisamente:

$$\hat{x}_{LS}(y) = \arg\min_{x\in\mathbb{R}^n} \underbrace{[y - h(x)]^T W [y - h(x)]}_{J_{LS}(x)} \tag{1.13}$$

dove $W = W^T > 0$ è una matrice di peso (simmetrica definita-positiva) opportunamente scelta. Si noti come la stima LS si riduca alla soluzione del problema di ottimizzazione, in

generale non lineare, (1.13) e, diversamente dai precedenti metodi di stima, non richieda alcuna ipotesi statistica/probabilistica né sulla variabile X da stimare né sull'errore di misura V. In altri termini, l'approccio LS è puramente deterministico. È interessante, tuttavia, verificare come si possa stabilire una connessione fra lo stimatore deterministico LS e lo stimatore ML nel caso in cui, per quest'ultimo, si ipotizzi un errore di misura Gaussiano $V = \mathcal{N}(0, \Sigma_V)$. In tal caso, infatti, la stima ML

$$\begin{aligned}
\hat{x}_{ML}(y) &= \arg\max_x f_{Y|X}(y|x) \\
&= \arg\max_x f_V(y - h(x)) \\
&= \arg\max_x \exp\left\{-[y - h(x)]^T \Sigma_V^{-1} [y - h(x)]\right\} \\
&= \arg\min_x [y - h(x)]^T \Sigma_V^{-1} [y - h(x)]
\end{aligned}$$

coincide con la stima LS (1.13) con matrice di peso $W = \Sigma_V^{-1}$ pari all'inversa della matrice di covarianza dell'errore di misura. A seguito di questi sviluppi, si può anche affermare che la stima LS con matrice di peso W può essere re-interpretata come stima ML nell'ipotesi di errore di misura Gaussiano a media nulla e di varianza $\Sigma_V = W^{-1}$.

È opportuno notare come la soluzione del problema di ottimizzazione (1.13) non sia, in generale, ottenibile analiticamente e debba piuttosto essere determinata numericamente, mediante metodi iterativi opportunamente inizializzati che, tuttavia, non sempre garantiscono convergenza al minimo globale della funzione $J_{LS}(x)$. Viceversa, nel caso particolare di modello di osservazione lineare, i.e. $h(X) = CX$, è facile mostrare che il problema (1.13) ammette soluzione analitica. Infatti, in questo caso, si ha

$$\begin{aligned}
\hat{x}_{LS}(y) &= \arg\min_x (y - Cx)^T W (y - Cx) \\
&= \arg\min_x \left\{x^T C^T W C x - 2x^T C^T W y + y^T W y\right\} \\
&= \arg\min_x \left\{x^T M x - 2x^T z + y^T W y\right\}
\end{aligned}$$

dove si è posto $M \triangleq C^T W C$ e $z \triangleq C^T W y$. Quindi, se la matrice $M = C^T W C \in \mathbb{R}^{n \times n}$ è invertibile, si ha

$$\begin{aligned}
\hat{x}_{LS}(y) &= \arg\min_x \left\{\left(x - M^{-1}z\right)^T M \left(x - M^{-1}z\right) + y^T W y - z^T M^{-1} z\right\} \\
&= M^{-1} z \\
&= \left(C^T W C\right)^{-1} C^T W y.
\end{aligned} \tag{1.14}$$

Si fa presente che la matrice $M = C^T W C$ risulta invertibile se e solo se $p \geq n$ e la matrice C ha n righe linearmente indipendenti. Nel caso di errore di misura $V = \mathcal{N}(0, \Sigma_V)$ Gaussiano, la formula (1.14), con $W = \Sigma_V^{-1}$, fornisce anche la stima di massima verosimiglianza, i.e., $\hat{x}_{ML} = \left(C^T \Sigma_V^{-1} C\right)^{-1} C^T \Sigma_V^{-1} y$.

Stima a minimo errore quadratico medio (MMSE)

Come indice di prestazione per valutare la qualità di uno stimatore si introduce l'*errore quadratico medio* (MSE = Mean Square Error).

Definizione di MSE - Dato lo stimatore $\widehat{X} = g(Y)$ se ne definisce l'MSE condizionato

$$\begin{aligned} MSE_g(y) &= E_{X|Y}\left[(X - g(Y))(X - g(Y))^T | Y = y\right] \\ &= \int (x - g(y))(x - g(y))^T \, f_{X|Y}(x|y)\, dx \end{aligned} \tag{1.15}$$

e l'MSE incondizionato

$$\begin{aligned} MSE_g &= E_{X,Y}\left[(X - g(Y))(X - g(Y))^T\right] \\ &= \int\int (x - g(y))(x - g(y))^T \, f_{X,Y}(x,y)\, dx\, dy \\ &= \int \left[\int (x - g(y))(x - g(y))^T \, f_{X|Y}(x|y)\, dx\right] f_Y(y)\, dy \\ &= \int MSE_g(y)\, f_Y(y)\, dy \\ &= E_Y\left\{E_{X|Y}\left[(X - g(Y))(X - g(Y))^T | Y = y\right]\right\}. \end{aligned} \tag{1.16}$$

Si noti come l'MSE, sia quello condizionato che quello non condizionato, sia una matrice $n \times n$, dove n è la dimensione della variabile X da stimare, simmetrica e semi-definita positiva. Poiché uno stimatore con MSE inferiore è certamente da preferirsi, viene naturale porsi il problema della determinazione dello stimatore a *minimo errore quadratico medio* (stimatore MMSE = Minimum Mean Squared Error) ovvero di uno stimatore $g^*(Y)$ il cui MSE sia inferiore a quello di ogni altro stimatore $g(Y)$. Per la precisione, considerando l'MSE condizionato, un tale stimatore deve soddisfare la disuguaglianza matriciale

$$E_{X|Y}\left[(X - g^*(Y))(X - g^*(Y))^T | Y\right] \leq E_{X|Y}\left[(X - g(Y))(X - g(Y))^T | Y\right], \ \forall g(\cdot) \tag{1.17}$$

dove, per matrici quadrate A e B della stessa dimensione, $A \leq B$ va interpretata nel senso che $A - B$ è semi-definita negativa, i.e. $A - B \leq 0$. Il seguente risultato fondamentale della teoria della stima sancisce che lo stimatore MMSE condizionato coincide con lo stimatore *media condizionata*.

Teorema della stima MMSE - Lo stimatore MMSE $g^*(\cdot)$ definito da (1.17) è dato da

$$g^*(Y) = E[X|Y] \tag{1.18}$$

e l'associato MMSE è dato da

$$\begin{aligned} E\left[(X-g^*(Y))\,(X-g^*(Y))^T\,|Y\right] &= var\,(X|Y) \\ &= E\left[XX^T|Y\right] - E\,[X|Y]\;\; E^T\,[X|Y]\,. \end{aligned} \tag{1.19}$$

Dimostrazione - Lo stimatore MMSE deve pertanto minimizzare, rispetto a tutti gli stimatori $g(\cdot)$, il funzionale di costo

$$\begin{aligned} V(g) &= E\left[(X-g(Y))\,(X-g(Y))^T\,|Y\right] \\ &= E\left[XX^T + g(Y)g^T(Y) - Xg^T(Y) - g(Y)X^T|Y\right] \\ &= E\left[XX^T|Y\right] + g(Y)g^T(Y) - E\,[X|Y]\,g^T(Y) - g(Y)E^T\,[X|Y] \\ &= \underbrace{(g(Y)-E\,[X|Y])\,(g(Y)-E\,[X|Y])^T}_{V_1(g)} + \underbrace{E\left[XX^T|Y\right] - E[X|Y]E^T[X|Y]}_{V_2} \end{aligned}$$

dove il primo termine $V_1(g)$, dipendente da $g(\cdot)$, può essere reso nullo (quindi minimo, essendo non-negativo) scegliendo $g(Y) = g^*(Y)$ uguale alla media condizionata come in (1.18), mentre il secondo termine V_2, indipendente da $g(\cdot)$, coincide con la varianza condizionata $var(X|Y) = E[XX^T|Y] - E[X|Y]E^T[X|Y] \geq 0$ e rappresenta il costo minimo (MMSE) $V(g^*)$. □

Pertanto lo stimatore MMSE coincide con lo stimatore Bayesiano *a media condizionata* (EAP) precedentemente introdotto. Di seguito, si mostra come tale stimatore risulti *non polarizzato*.

Teorema 1 (Non polarizzazione dello stimatore MMSE) - Lo stimatore MMSE $g^*(Y) = E\,[X|Y]$ è non polarizzato nel senso che

$$E_Y\,[g^*(Y)] \;=\; E_X\,[X]\,.$$

Dimostrazione - Si ha infatti:

$$
\begin{aligned}
E_Y\left[g^*(Y)\right] &= \int g^*(y) f_Y(y) dy \\
&= \int E[X|Y=y] f_Y(y) dy \\
&= \int \left[\int x f_{X|Y}(x|y)\, dx \right] f_Y(y) dy \\
&= \int\int x f_{X,Y}(x,y) dx\, dy \\
&= \int x \left[\int f_{X,Y}(x,y)\, dy \right] dx \\
&= \int x f_X(x) dx \\
&= E_X\left[X\right]
\end{aligned}
$$

come volevasi dimostrare. □

Benchè lo stimatore MMSE sia stato ottenuto minimizzando l'MSE condizionato (1.15), si mostra di seguito come il medesimo stimatore minimizzi anche l'MSE incondizionato (1.16).

Teorema 2 (Ottimalità incondizionata dello stimatore MMSE) - Lo stimatore MMSE $g^*(Y) = E\left[X|Y\right]$ minimizza anche l'MSE non condizionato nel senso che

$$
E_{X,Y}\left[\left(X - g^*(Y)\right)\left(X - g^*(Y)\right)^T\right] \;\leq\; E_{X,Y}\left[\left(X - g(Y)\right)\left(X - g(Y)\right)^T\right], \quad \forall\, g(\cdot) \tag{1.20}
$$

Dimostrazione - Il risultato si basa sulla seguente proprietà dell'operatore di media:

$$
E_Y\left\{E_{X|Y}\left[\psi(X,Y)|Y=y\right]\right\} \;=\; E_{X,Y}\left[\psi(X,Y)\right] \tag{1.21}
$$

dove $\psi(\cdot,\cdot)$ è una arbitraria funzione delle variabili aleatorie X e Y.
Infatti,

$$
\begin{aligned}
E_{X,Y}\left[\psi(X,Y)\right] &= \int\int \psi(x,y) f_{X,Y}(x,y)\, dx\, dy \\
&= \int\int \psi(x,y) f_{X|Y}(x|y) f_Y(y)\, dx\, dy \\
&= \int \left[\int \psi(x,y) f_{X|Y}(x|y)\, dx \right] f_Y(y)\, dy \\
&= \int E_{X|Y}\left[\psi(X,Y)|Y=y\right] f_Y(y)\, dy \\
&= E_Y\left\{E_{X|Y}\left[\psi(X,Y)|Y=y\right]\right\}.
\end{aligned}
$$

Per definizione, lo stimatore MMSE $g^*(Y) = E[X|Y]$ soddisfa (1.17). Applicando ad ambo i membri di (1.17) l'operatore E_Y e sfruttando la proprietà (1.21), si ottiene quindi la relazione (1.20) come volevasi dimostrare. □

Teorema 3 (Principio di ortogonalità della stima MMSE) - L'errore di stima dello stimatore MMSE è ortogonale a (incorrelato con) tutti gli stimatori $g(Y)$, i.e.,

$$E\left\{\left(X - E\left[X|Y\right]\right) g^T(Y)|Y\right\} \;=\; 0, \qquad \forall g(\cdot). \tag{1.22}$$

Dimostrazione - Si ha banalmente

$$E\left\{\left(X - E\left[X|Y\right]\right) g^T(Y)|Y\right\} \;=\; E\left[X|Y\right] g^T(Y) - E\left[X|Y\right] g^T(Y) \;=\; 0$$

qualunque sia lo stimatore $g(\cdot)$, come volevasi dimostrare. □

In altri termini, la stima MMSE di X basata su Y risulta la proiezione ortogonale della variabile X sullo spazio degli stimatori, ovvero delle funzioni $g(Y)$.

Teorema 4 (Stima MMSE di funzioni affini del parametro X) - Sia $g^*(Y) = E[X|Y]$ lo stimatore MMSE di X basato sull'osservazione Y e siano $A \in \mathbb{R}^{m\times n}$ e $b \in \mathbb{R}^m$. Allora $\gamma^*(Y) = Ag^*(Y) + b$ è lo stimatore MMSE di $Z \stackrel{\triangle}{=} AX + b \in \mathbb{R}^m$ basato su Y.

Dimostrazione - In virtù della linearità dell'operatore di media: $\gamma^*(Y) = E\left[Z|Y\right] = E\left[AX + b|Y\right] = AE[X|Y] + b = Ag^*(Y) + b$, qualunque siano A e b di dimensioni opportune, come volevasi dimostrare. □

Teorema 5 (Stima MMSE nel caso Gaussiano) - Se X e Y sono congiuntamente Gaussiane, i.e.,

$$\left[\begin{array}{c} X \\ Y \end{array}\right] \;=\; \mathcal{N}\left(\left[\begin{array}{c} \overline{x} \\ \overline{y} \end{array}\right], \left[\begin{array}{cc} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{array}\right]\right), \quad \Sigma_{YX} = \Sigma_{XY}^T,$$

allora lo stimatore MMSE è dato da

$$\begin{aligned} \widehat{X} = g^*(Y) = E\left[X|Y\right] &= A^*Y + b^* \\ A^* &= \Sigma_{XY}\Sigma_Y^{-1} \\ b^* &= \overline{x} - A^*\overline{y} = \overline{x} - \Sigma_{XY}\Sigma_Y^{-1}\overline{y} \end{aligned} \tag{1.23}$$

con relativo MMSE dato da

$$\widehat{\Sigma}_X \stackrel{\triangle}{=} E\left[\left(X - g^*(Y)\right)\left(X - g^*(Y)\right)^T\right] \;=\; \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T \tag{1.24}$$

Dimostrazione - Nelle ipotesi fatte, sfruttando le formule della media e covarianza condizionata di variabili aleatorie Gaussiane, si ha

$$\begin{aligned}\widehat{X} &= g^*(Y) = E[X|Y] \\ &= \overline{x} + \Sigma_{XY}\Sigma_Y^{-1}(Y - \overline{y}) \\ &= \Sigma_{XY}\Sigma_Y^{-1}Y + \left[\overline{x} - \Sigma_{XY}\Sigma_Y^{-1}\overline{y}\right] \\ &= A^*Y + b^*.\end{aligned}$$

Inoltre,

$$\widehat{\Sigma}_X = var(X|Y) = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T$$

come volevasi dimostrare. □

Stima lineare ottima (BLUE)

Definizione di stimatore affine - Uno stimatore della forma $g(Y) = AY + b$, per opportuni $A \in \mathbb{R}^{n\times p}$ e $b \in \mathbb{R}^n$, dicesi *stimatore affine* (*lineare* se $b = 0$) di X basato su Y.

Pertanto il precedente risultato afferma che, nel caso Gaussiano, lo stimatore MMSE risulta affine. Dal momento che uno stimatore affine, completamente caratterizzato dalla matrice $A \in \mathbb{R}^{n\times p}$ e dal vettore $b \in \mathbb{R}^n$, è di facile implementazione, viene naturale porsi il problema di determinare il migliore, nel senso del minimo errore quadratico medio, stimatore affine $a^*(Y) = A^*Y + b^*$. In altri termini, fra tutti gli stimatori affini $a(Y) = AY + b$ si cerca quello che:

(1) fornisce polarizzazione nulla, i.e.,

$$E\left[\widetilde{X}\right] = E\left[X - a(Y)\right] = E\left[X - AY - b\right] = \overline{x} - A\overline{y} - b = 0$$

da cui

$$b = \overline{x} - A\overline{y}; \tag{1.25}$$

(2) rende minimo l'MSE

$$\begin{aligned} V(a) &= E\left[(X-a(Y))\,(X-a(Y))^T\right] \\ &= E\left[(X-AY-b)\,(X-AY-b)^T\right] \\ &= E\left[(X-AY-\overline{x}+A\overline{y})\,(X-AY-\overline{x}+A\overline{y})^T\right] \\ &= E\left[\left(\widetilde{X}-A\widetilde{Y}\right)\left(\widetilde{X}-A\widetilde{Y}\right)^T\right] \\ &= E\left[\widetilde{X}\widetilde{X}^T\right]-AE\left[\widetilde{Y}\widetilde{X}^T\right]-E\left[\widetilde{X}\widetilde{Y}^T\right]A^T+AE\left[\widetilde{Y}\widetilde{Y}^T\right]A^T \\ &= \Sigma_X-A\Sigma_{YX}-\Sigma_{XY}A^T+A\Sigma_Y A^T \\ &= \Sigma_X-A\Sigma_{XY}^T-\Sigma_{XY}A^T+A\Sigma_Y A^T \\ &= \left(A-\Sigma_{XY}\Sigma_Y^{-1}\right)\Sigma_Y\left(A-\Sigma_{XY}\Sigma_Y^{-1}\right)^T+\left[\Sigma_X-\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T\right] \end{aligned} \tag{1.26}$$

Lo stimatore in oggetto viene riferito in letteratura come *migliore stimatore lineare non polarizzato* (*BLUE = Best Linear Unbiased Estimator*), seppur con una imprecisione terminologica trattandosi, in generale, di uno stimatore affine. Vale il seguente risultato.

Teorema della stima BLUE - Sia

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \left(\begin{bmatrix} \overline{x} \\ \overline{y} \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_Y \end{bmatrix}\right)$$

Allora lo stimatore BLUE di X basato su Y è dato da $a^*(Y)=A^*Y+b^*$ con

$$\begin{cases} A^* &= \Sigma_{XY}\Sigma_Y^{-1} \\ b^* &= \overline{x}-A^*\overline{y} \;=\; \overline{x}-\Sigma_{XY}\Sigma_Y^{-1}\overline{y} \end{cases} \tag{1.27}$$

ed il relativo MSE risulta

$$\widehat{\Sigma}_X \stackrel{\triangle}{=} E\left[(X-a^*(Y))\,(X-a^*(Y))^T\right] \;=\; \Sigma_X-\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T. \tag{1.28}$$

Dimostrazione - Da (1.26) si evince immediatamente che la miglior scelta della matrice A dello stimatore affine è proprio $A=A^*\stackrel{\triangle}{=}\Sigma_{XY}\Sigma_Y^{-1}$. Conseguentemente, dal vincolo di non polarizzazione dello stimatore (1.25), si ricava $b=b^*=\overline{x}-A^*\overline{y}=\overline{x}-\Sigma_{XY}\Sigma_Y^{-1}\overline{y}$. Infine, sostituendo $A=A^*=\Sigma_{XY}\Sigma_Y^{-1}$ in (1.26), si ottiene $\widehat{\Sigma}_X\stackrel{\triangle}{=}V(a^*)=\Sigma_X-\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T$ come volevasi dimostrare. □

Nome	**Stimatore** $\widehat{X}(Y)$	**Stima** $\hat{x}(y)$
EAP	$E[X\|Y]$	$E[X\|Y=y]$
MAP		$\arg\max_x f_{Y\|X}(y\|x)f_X(x)$
ML		$\arg\max_x f_{Y\|X}(y\|x)$
WLS		$\arg\min_x (y-h(x))^T W(y-h(x))$
MMSE	$g^*(Y) = \arg\min_g E\left[(X-g(Y))(X-g(Y))^T\right]$ $g^*(Y) = E[X\|Y]$	$g^*(y) = \frac{\int x f_{Y\|X}(y\|x)f_X(x)dx}{\int f_{Y\|X}(y\|x)f_X(x)dx}$
BLUE	$a^*(Y) = A^*Y + b^*$ $(A^*, b^*) = \arg\min_{A,b} E\left[(X-AY-b)(X-AY-b)^T\right]$ subject to $AE[Y]+b = E[X]$ $A^* = \Sigma_{XY}\Sigma_Y^{-1}, b^* = \overline{x} - \Sigma_{XY}\Sigma_Y^{-1}\overline{y}$	$a^*(y) = \overline{x} + \Sigma_{XY}\Sigma_Y^{-1}(y-\overline{y})$

Tabella 1.1: Tabella riassuntiva degli stimatori presentati.

Nome	**Equivalenze**
EAP	coincide sempre con MMSE
MAP	coincide con ML se non si ha informazione a priori su X coincide con EAP $\equiv$ MMSE se X & Y sono congiuntamente Gaussiane
ML	coincide con WLS se $V = \mathcal{N}(0, W^{-1})$
WLS	coincide con ML nel caso $V = \mathcal{N}(0, W^{-1})$
MMSE	coincide sempre con EAP
BLUE	coincide con MMSE nel caso di modello di osservazione lineare-Gaussiano

Tabella 1.2: Relazioni fra i vari stimatori.

Si noti come nel caso Gaussiano gli stimatori MMSE e BLUE coincidano mentre in generale lo stimatore MMSE risulta non lineare e, di conseguenza, lo stimatore BLUE è soltanto sub-ottimo in senso assoluto, sebbene sia ottimo fra tutti gli stimatori affini non polarizzati. La tabella 1.1 riassume gli stimatori (EAP, MAP, ML, WLS, MMSE, BLUE) considerati mentre la tabella 1.2 riassume le relazioni esistenti fra i vari stimatori.

L'esempio successivo vuole evidenziare, in un problema di stima lineare ma non Gaussiano, la diversità degli stimatori BLUE ed MMSE.

Esempio (Stima di variabile affetta da errore di misura uniforme) - Si consideri il problema della stima di una grandezza scalare X da p osservazioni rumorose

$$Y_i \;=\; X + V_i \qquad\qquad i = 1, \cdots, p$$

con errori di misura $V_1, \ldots, V_p$ indipendenti ed identicamente distribuiti, nonché indipen-

denti da X. Il modello di osservazione lineare risultante è

$$\begin{cases} Y \stackrel{\triangle}{=} [Y_1, \ldots, Y_p]^T &= CX + V \\ X &\sim (\overline{x}, \sigma_X^2), \quad \sigma_X^2 > 0 \\ V &\sim (0, \sigma_V^2 I_p), \quad \sigma_V^2 > 0 \\ C &= \mathbf{1_p} \stackrel{\triangle}{=} [1, \ldots, 1]^T \in \mathbb{R}^p. \end{cases} \tag{1.29}$$

Si assume che X abbia distribuzione Gaussiana e che, viceversa, ciascun errore di misura V_i sia uniformemente distribuito nell'intervallo $[-\delta, \delta]$, i.e. a media nulla e di varianza $\sigma_V^2 = \delta^2/3$. Pertanto, si ha:

$$\begin{aligned} X = \mathcal{N}(\overline{x}, \sigma_X^2) \;\Rightarrow\; f_X(x) &\propto \exp\left[-\frac{(x-\overline{x})^2}{2\sigma_X^2}\right] \\ V = \mathcal{U}([-\delta, \delta]^p) \;\Rightarrow\; f_V(v) = f_V(v_1, \ldots, v_p) &\propto \prod_{i=1}^{n} [1(v_i + \delta) - 1(v_i - \delta)] \end{aligned}$$

dove $1(\cdot)$ denota la funzione a gradino unitario definita come segue

$$1(x) \;=\; \begin{cases} 1, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Poiché l'errore di misura V non è Gaussiano, $Y = \mathbf{1_p}X + V$ non risulta a sua volta Gaussiano; pertanto X e Y non sono congiuntamente Gaussiane per cui ci si aspetta uno stimatore MMSE non affine e, di conseguenza, diverso dallo stimatore BLUE. Nel seguito si procede alla determinazione analitica dello stimatore BLUE e, successivamente, dello stimatore MMSE in accordo alla teoria precedentemente esposta.

Stimatore BLUE - È facile verificare che, in riferimento al modello di osservazione (1.29), si ha

$$\begin{cases} \overline{y} &= C\overline{x} &&= \mathbf{1_p}\overline{x} \\ \Sigma_{XY} &= E\left[\widetilde{X}\widetilde{Y}^T\right] = E\left[\widetilde{X}^2\right] C^T = \sigma_X^2 C^T &&= \sigma_X^2 \mathbf{1_p^T} \\ \Sigma_Y &= E\left[\widetilde{Y}\widetilde{Y}^T\right] = E\left[\left(C\widetilde{X} + V\right)\left(C\widetilde{X} + V\right)^T\right] &&= \mathbf{1_p}\,\sigma_X^2 \mathbf{1_p^T} + \sigma_V^2 I_p \end{cases}$$

per cui, dal teorema della stima BLUE, si ha

$$\begin{aligned}\hat{x}_{BLUE} &= \overline{x} + \underbrace{\sigma_X^2 \mathbf{1_p^T}}_{\Sigma_{XY}} \underbrace{\left(\sigma_V^2 I_p + \mathbf{1_p}\sigma_X^2 \mathbf{1_p^T}\right)^{-1}}_{\Sigma_Y^{-1}} \left(y - \underbrace{\mathbf{1_p}\overline{x}}_{\overline{y}}\right) \\ &= \left[\sigma_X^2 - \sigma_X^2 \mathbf{1_p^T}\left(\sigma_V^2 I_p + \mathbf{1_p}\sigma_X^2 \mathbf{1_p^T}\right)^{-1} \mathbf{1_p}\sigma_X^2\right] \frac{\overline{x}}{\sigma_X^2} + \sigma_X^2 \mathbf{1_p^T}\left(\sigma_V^2 I_p + \mathbf{1_p}\sigma_X^2 \mathbf{1_p^T}\right)^{-1} y \end{aligned} \tag{1.30}$$

con associato MSE

$$\hat{\sigma}_X^2 \stackrel{\triangle}{=} E\left[(X - \hat{x}_{BLUE})^2\right] = \sigma_X^2 - \sigma_X^2 \mathbf{1_p^T}\left(\sigma_V^2 I_p + \mathbf{1_p}\sigma_X^2 \mathbf{1_p^T}\right)^{-1} \mathbf{1_p}\sigma_X^2 \tag{1.31}$$

Per i successivi sviluppi di (1.30) e (1.31) conviene introdurre i seguenti due risultati di algebra delle matrici.

Lemma di inversione di matrice - Siano P e Q due matrici quadrate ed invertibili. Allora:

$$\left(P + CQC^T\right)^{-1} = P^{-1} - P^{-1}C\left(Q^{-1} + C^T P^{-1} C\right)^{-1} C^T P^{-1}. \tag{1.32}$$

Dimostrazione - Si procede per verifica diretta

$$\begin{aligned}&\left(P + CQC^T\right)\left[P^{-1} - P^{-1}C\left(Q^{-1} + C^T P^{-1}C\right)^{-1} C^T P^{-1}\right] \\ &= I - C\left(Q^{-1} + C^T P^{-1}C\right)^{-1} C^T P^{-1} + CQC^T P^{-1} - CQC^T P^{-1} C\left(Q^{-1} + C^T P^{-1}C\right)^{-1} C^T P^{-1} \\ &= I + CQC^T P^{-1} - C\left(I + QC^T P^{-1}C\right)\left(Q^{-1} + C^T P^{-1}C\right) C^T P^{-1} \\ &= I + CQC^T P^{-1} - CQ\left(Q^{-1} + C^T P^{-1}C\right)\left(Q^{-1} + C^T P^{-1}C\right)^{-1} C^T P^{-1} \\ &= I + CQC^T P^{-1} - CQC^T P^{-1} \\ &= I\end{aligned}$$

come volevasi dimostrare. □

Il seguente risultato è una immediata conseguenza del precedente lemma di inversione di matrice.

Corollario del lemma di inversione di matrice - Se P e Q sono matrici invertibili, allora

$$QC^T\left(P + CQC^T\right)^{-1} = \left(Q^{-1} + C^T P^{-1}C\right)^{-1} C^T P^{-1} \tag{1.33}$$

Dimostrazione - Applicando al primo membro di (1.33) il lemma di inversione di matrice, si ha

$$\begin{aligned} QC^T\left(P+CQC^T\right)^{-1} &= QC^T\left[P^{-1}-P^{-1}C\left(Q^{-1}+C^TP^{-1}C\right)^{-1}C^TP^{-1}\right] \\ &= QC^TP^{-1}-QC^TP^{-1}C\left(Q^{-1}+C^TP^{-1}C\right)^{-1}C^TP^{-1} \\ &= Q\left[I-C^TP^{-1}C\left(Q^{-1}+C^TP^{-1}C\right)^{-1}\right]C^TP^{-1} \\ &= Q\left(Q^{-1}+C^TP^{-1}C-C^TP^{-1}C\right)\left(Q^{-1}+C^TP^{-1}C\right)^{-1}C^TP^{-1} \\ &= \left(Q^{-1}+C^TP^{-1}C\right)^{-1}C^TP^{-1} \end{aligned}$$

come volevasi dimostrare. □

Applicando il lemma di inversione di matrice a (1.31) con $P = 1/\sigma_X^2$, $Q = 1/\sigma_V^2$ e $C = \mathbf{1_p}$ si ottiene:

$$\hat{\sigma}_X^2 = \left(\frac{1}{\sigma_X^2}+\frac{\mathbf{1_p^T}\mathbf{1_p}}{\sigma_V^2}\right)^{-1} = \left(\frac{1}{\sigma_X^2}+\frac{p}{\sigma_V^2}\right)^{-1} = \frac{\sigma_V^2\sigma_X^2}{\sigma_V^2+p\sigma_X^2} \tag{1.34}$$

Quindi, applicando (1.33) a (1.30) con $Q = \sigma_X^2$, $C = \mathbf{1_p}$ e $P = \sigma_V^2 I_p$ si ottiene

$$\begin{aligned} \hat{x}_{BLUE} &= \hat{\sigma}_X^2\frac{\overline{x}}{\sigma_X^2}+\left(\frac{1}{\sigma_X^2}+\frac{p}{\sigma_V^2}\right)^{-1}\frac{\mathbf{1_p^T}y}{\sigma_V^2} = \hat{\sigma}_X^2\left(\frac{\overline{x}}{\sigma_X^2}+\frac{\mathbf{1_p^T}y}{\sigma_V^2}\right) \\ &= \frac{\sigma_V^2\sigma_X^2}{\sigma_V^2+p\sigma_X^2}\left(\frac{\overline{x}}{\sigma_X^2}+\frac{\mathbf{1_p^T}y}{\sigma_V^2}\right) \\ &= \frac{\sigma_V^2}{\sigma_V^2+p\sigma_X^2}\overline{x}+\frac{p\sigma_X^2}{\sigma_V^2+p\sigma_X^2}\frac{\mathbf{1_p^T}y}{p} \\ &= \frac{\sigma_V^2}{\sigma_V^2+p\sigma_X^2}\overline{x}+\frac{p\sigma_X^2}{\sigma_V^2+p\sigma_X^2}\left(\frac{1}{p}\sum_{i=1}^{p}y_i\right) \\ &= \lambda\overline{x}+(1-\lambda)\check{y} \end{aligned} \tag{1.35}$$

con

$$\lambda \stackrel{\triangle}{=} \frac{\sigma_V^2}{\sigma_V^2+p\sigma_X^2} \in (0,1), \qquad \check{y} \stackrel{\triangle}{=} \frac{1}{p}\sum_{i=1}^{p}y_i.$$

Si noti che $\check{y}$ rappresenta la media aritmetica (campionaria) delle osservazioni disponibili. Quindi la stima BLUE (1.35) risulta combinazione lineare convessa della media $\overline{x}$ (stima

a priori) di X e della media campionaria $\breve{y}$ delle misure. Il corrispondente MSE dello stimatore BLUE è dato da (1.34) e può essere espresso equivalentemente nella forma

$$\hat{\sigma}_X^2 = \frac{\frac{\sigma_V^2}{p}\,\sigma_X^2}{\frac{\sigma_V^2}{p} + \sigma_X^2} = \frac{\sigma_V^2}{p} \parallel \sigma_X^2 \;\leq\; \min\left(\frac{\sigma_V^2}{p}, \sigma_X^2\right) \tag{1.36}$$

dove $a \parallel b \triangleq ab(a+b)^{-1}$ denota il *parallelo* di a e b, formula utilizzata ad esempio per il calcolo della resistenza/impedenza equivalente a due resistenze/impedenze in parallelo a e b. Pertanto (1.36) esprime il fatto (*legge del parallelo*) che la varianza a posteriori $\hat{\sigma}_X^2$ è il parallelo della varianza a priori σ_X^2 e di σ_V^2/p, la varianza σ_V^2 di ogni singolo errore di misura divisa per il numero p di misure indipendenti. È evidente che $\hat{\sigma}_X^2$ è una funzione monotòna decrescente di p, cioè la varianza a posteriori si riduce all'aumentare del numero di osservazioni come era lecito attendersi, e che questa converge a zero quando il numero di osservazioni cresce indefinitamente, i.e. $\lim_{p\to\infty} \hat{\sigma}_X^2 = 0$.

Si noti che, nel caso limite in cui non si abbia informazione a priori su X, i.e. quando $\sigma_X^2/\sigma_V^2 \to \infty$, risulta che

$$\hat{x}_{BLUE} \to \breve{y} \triangleq \frac{1}{p}\sum_{i=1}^{p} y_i, \quad \hat{\sigma}_X^2 \to \frac{\sigma_V^2}{p}$$

ovvero la stima BLUE tende alla media aritmetica delle misure con MSE che decresce all'aumentare del numero di misure p come $1/p$. Viceversa, nel caso limite opposto in cui $\sigma_X^2/\sigma_V^2 \to 0$ (osservazioni non informative), si ha

$$\hat{x}_{BLUE} \to \overline{x}, \quad \hat{\sigma}_X^2 \to \sigma_X^2.$$

In qualunque situazione intermedia fra i due casi limite sopra considerati, risulta che

$$\hat{x}_{BLUE} \in (\min\{\overline{x}, \breve{y}\}, \max\{\overline{x}, \breve{y}\}), \quad \hat{\sigma}_X^2 < \min\left\{\sigma_X^2, \frac{\sigma_V^2}{p}\right\}$$

ovvero la stima BLUE si trova all'interno del segmento che congiunge le stime $\overline{x}$ e $\breve{y}$ relative ai due casi limite, ed il suo MSE è inferiore ad entrambi i valori σ_X^2 e σ_V^2/p che si otterebbero nei casi limite di osservazioni non informative e, rispettivamente, nessuna informazione a priori; questo conferma che la stima BLUE combina sempre in modo efficace le due sorgenti di informazione, informazione a priori ed osservazioni, riducendo comunque l'incertezza rispetto all'uso di una sola delle due sorgenti.

Stimatore MMSE - La stima MMSE coincide con la media condizionata

$$\hat{x}_{MMSE}(y) = E[X|Y=y] = \frac{\int x f_{Y|X}(y|x) f_X(x) dx}{\int f_{Y|X}(y|x) f_X(x) dx}$$

dove

$$\begin{aligned} f_{Y|X}(y|x) &= f_V\left(y - \mathbf{1_p} x\right) \\ &= \frac{1}{(2\delta)^p} \prod_{i=1}^{p} \left[1\left(y_i - x + \delta\right) - 1\left(y_i - x - \delta\right)\right] \\ &= \frac{1}{(2\delta)^p} \left[1\left(y_{max} - \delta - x\right) - 1\left(y_{min} + \delta - x\right)\right] \end{aligned}$$

con

$$y_{min} \stackrel{\triangle}{=} \min_{1 \le i \le p} y_i, \qquad y_{max} \stackrel{\triangle}{=} \max_{1 \le i \le p} y_i.$$

Pertanto, posto $a \stackrel{\triangle}{=} y_{max} - \delta$ e $b \stackrel{\triangle}{=} y_{min} + \delta$, si ha

$$\begin{aligned} \hat{x}_{MMSE}(y) &= \frac{\displaystyle\int_a^b x \exp\left[-\frac{(x-\overline{x})^2}{2\sigma_X^2}\right] dx}{\displaystyle\int_a^b \exp\left[-\frac{(x-\overline{x})^2}{2\sigma_X^2}\right] dx} = \overline{x} + \frac{\displaystyle\int_a^b (x-\overline{x}) \exp\left[-\frac{(x-\overline{x})^2}{2\sigma_X^2}\right] dx}{\displaystyle\int_a^b \exp\left[-\frac{(x-\overline{x})^2}{2\sigma_X^2}\right] dx} \\ &= \overline{x} + \frac{\sigma_X}{\sqrt{2\pi}} \frac{\exp\left[-\frac{(a-\overline{x})^2}{2\sigma_X^2}\right] - \exp\left[-\frac{(b-\overline{x})^2}{2\sigma_X^2}\right]}{Q\left(\frac{a-\overline{x}}{\sigma_X}\right) - Q\left(\frac{b-\overline{x}}{\sigma_X}\right)} \\ &= \overline{x} + \frac{\sigma_X}{\sqrt{2\pi}} \frac{\exp\left[-\frac{(y_{max}-\delta-\overline{x})^2}{2\sigma_X^2}\right] - \exp\left[-\frac{(y_{min}+\delta-\overline{x})^2}{2\sigma_X^2}\right]}{Q\left(\frac{y_{max}-\delta-\overline{x}}{\sigma_X}\right) - Q\left(\frac{y_{min}+\delta-\overline{x}}{\sigma_X}\right)} \end{aligned} \tag{1.37}$$

dove la funzione $Q(\cdot)$ è definita come segue

$$Q(z) \stackrel{\triangle}{=} \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\zeta^2/2} d\zeta.$$

Si noti come la stima MMSE fornita da (1.37) dipenda soltanto, in modo non lineare, dalle due misure estreme $y_{min} \stackrel{\triangle}{=} \min_i y_i$ e $y_{max} \stackrel{\triangle}{=} \max_i y_i$ e, nel caso $p > 2$, ignori le altre $p-2$ misure intermedie. Questo è in contrasto con la stima BLUE che dipende linearmente da tutte le misure $\{y_i\}_{i=1}^p$ mediante la loro media aritmetica. In particolare, nel caso di assenza di informazione a priori i.e. quando $\sigma_X \to \infty$, si può considerare la PDF a priori

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_X} \exp\left[-\frac{(x-\overline{x})^2}{2\sigma_X^2}\right]$$

pressoché costante nell'intervallo di integrazione $[a,b] = [y_{max} - \delta, y_{min} + \delta]$ e pertanto

$$\begin{aligned}\hat{x}_{MMSE}(y) &= \frac{\displaystyle\int_a^b x \exp\left[-\frac{(x-\overline{x})^2}{2\sigma_X^2}\right] dx}{\displaystyle\int_a^b \exp\left[-\frac{(x-\overline{x})^2}{2\sigma_X^2}\right] dx} \overbrace{\longrightarrow}^{\sigma_X \to \infty} \frac{\displaystyle\int_a^b x\, dx}{\displaystyle\int_a^b dx} \\ &= \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2} = \frac{y_{min} + y_{max}}{2}\end{aligned}$$

In altri termini lo stimatore MMSE, in assenza di informazione a priori, effettua la media aritmetica delle due misure estreme in contrapposizione allo stimatore BLUE che media tutte le p misure disponibili.

Si dimostra che l'MSE associato allo stimatore $\hat{x}_{MMSE}(y) = (y_{min} + y_{max})/2$ è dato da

$$\hat{\sigma}_X^2 = \frac{6\sigma_V^2}{(p+1)(p+2)} \tag{1.38}$$

da confrontare con l'MSE $\hat{\sigma}_X^2 = \sigma_V^2/p$ dello stimatore $\hat{x}_{BLUE}(y) = p^{-1}\displaystyle\sum_{i=1}^{p} y_i$.

Stima nel caso di modelli di osservazione lineare

In questa sezione si studia in dettaglio il problema della stima parametrica nel caso di modello di osservazione lineare

$$\begin{cases} Y = CX + V \\ X \sim (\overline{x}, \Sigma_X)\,,\ \Sigma_X > 0 \\ V \sim (0, \Sigma_V)\,,\ \Sigma_V > 0 \\ X \perp V. \end{cases} \tag{1.39}$$

È immediato constatare che:

$$\begin{aligned}\overline{y} &\triangleq E[Y] = E[CX + V] = CE[X] + E[V] = C\overline{x} \\ \widetilde{Y} &\triangleq Y - \overline{y} = CX + V - C\overline{x} = C(X - \overline{x}) + V = C\widetilde{X} + V \\ \Sigma_{XY} &\triangleq E\left[\widetilde{X}\widetilde{Y}^T\right] = E\left[\widetilde{X}\left(C\widetilde{X} + V\right)^T\right] = E\left[\widetilde{X}\widetilde{X}^T\right]C^T + E\left[\widetilde{X}V^T\right] = \Sigma_X C^T \\ \Sigma_Y &\triangleq E\left[\widetilde{Y}\widetilde{Y}^T\right] = E\left[\left(C\widetilde{X} + V\right)\left(C\widetilde{X} + V\right)^T\right] = CE\left[\widetilde{X}\widetilde{X}^T\right]C^T + E[VV^T] \\ &= C\Sigma_X C^T + \Sigma_V > 0\end{aligned}$$

da cui si ottiene la stima BLUE di X basata su Y

$$\begin{aligned}
\hat{x}(y) = \hat{x} &= \overline{x} + \Sigma_{XY}\Sigma_Y^{-1}(y - \overline{y}) \\
&= \overline{x} + \Sigma_X C^T \left(\Sigma_V + C\Sigma_X C^T\right)^{-1}(y - C\overline{x}) \\
&= \left[I - \Sigma_X C^T \left(\Sigma_V + C\Sigma_X C^T\right)^{-1} C\right]\overline{x} + \Sigma_X C^T \left(\Sigma_V + C\Sigma_X C^T\right)^{-1} y \\
&= \left[\Sigma_X - \Sigma_X C^T \left(\Sigma_V + C\Sigma_X C^T\right)^{-1} C\Sigma_X\right]\Sigma_X^{-1}\overline{x} + \Sigma_X C^T \left(\Sigma_V + C\Sigma_X C^T\right)^{-1} y \\
&= \left(\Sigma_X^{-1} + C^T\Sigma_V^{-1}C\right)^{-1}\Sigma_X^{-1}\overline{x} + \left(\Sigma_X^{-1} + C^T\Sigma_V^{-1}C\right)^{-1} C^T\Sigma_V^{-1} y \\
&= \left(\Sigma_X^{-1} + C^T\Sigma_V^{-1}C\right)^{-1}\ \left(\Sigma_X^{-1}\overline{x} + C^T\Sigma_V^{-1} y\right)
\end{aligned} \tag{1.40}$$

dove nel penultimo passaggio sono state utilizzate le formule (1.32) e (1.33). L'MSE relativo a tale stima è dato da

$$\begin{aligned}
\widehat{\Sigma}_X &\triangleq E\left[(X - \hat{x})(X - \hat{x})^T\right] \\
&= \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T \\
&= \Sigma_X - \Sigma_X C^T \left(\Sigma_V + C\Sigma_X C^T\right)^{-1} C\Sigma_X \\
&= \left(\Sigma_X^{-1} + C^T\Sigma_V^{-1}C\right)^{-1}
\end{aligned} \tag{1.41}$$

da cui

$$\begin{cases}
\widehat{\Sigma}_X^{-1} &= \Sigma_X^{-1} + C^T\Sigma_V^{-1}C \\
\widehat{\Sigma}_X^{-1}\hat{x} &= \Sigma_X^{-1}\overline{x} + C^T\Sigma_V^{-1} y.
\end{cases} \tag{1.42}$$

Le due formule in (1.42) forniscono una rappresentazione compatta e di immediata interpretazione della soluzione del problema di stima parametrica lineare. In primo luogo si ricorda che la stima $\hat{x}$ coincide con la stima MMSE, i.e. la media condizionata $E[X|Y = y]$, nel caso Gaussiano cioè quando $X = \mathcal{N}(\overline{x}, \Sigma_X)$ e $V = \mathcal{N}(0, \Sigma_V)$. Da (1.42) si comprende chiaramente quello che accade nei casi limite di *assenza di informazione a priori* ($\Sigma_X^{-1} = 0$) e di *assenza di contributo informativo da parte delle misure* ($\Sigma_V^{-1} = 0$). Infatti nel primo caso,

$$\hat{x} \to \left(C^T\Sigma_V^{-1}C\right)^{-1}\ C^T\Sigma_V^{-1} y, \quad \widehat{\Sigma}_X \to \left(C^T\Sigma_V^{-1}C\right)^{-1}\ \text{per } \Sigma_X^{-1} \to 0$$

cioè la stima tende alla stima ai minimi quadrati con matrice di peso $W = \Sigma_V^{-1}$ che, a sua volta, nel caso di errore di misura V Gaussiano, coincide con la stima di massima verosimiglianza. Viceversa, nel secondo caso,

$$\hat{x} \to \overline{x}, \ \ \widehat{\Sigma}_X \to \Sigma_X \text{ per } \Sigma_V^{-1} \to 0$$

cioè stima e covarianza a-posteriori tendono ai loro corrispettivi a-priori.

Per gli sviluppi successivi si osserva che la coppia (stima, covarianza), sia quella a-priori $(\overline{x}, \Sigma_X)$ che quella a-posteriori $\left(\hat{x}, \widehat{\Sigma}_X\right)$, può essere equivalentemente rappresentata dalla *coppia di informazione* $(q_X, \Omega_X) \triangleq \left(\Sigma_X^{-1}\overline{x}, \Sigma_X^{-1}\right)$ o rispettivamente $\left(\hat{q}_X, \widehat{\Omega}_X\right) \triangleq \left(\widehat{\Sigma}_X^{-1}\hat{x}, \widehat{\Sigma}_X^{-1}\right)$. Si noti che:

- la matrice di informazione Ω è l'inversa della matrice di covarianza Σ, mentre il vettore di informazione q è il prodotto della matrice di informazione per il vettore stima;
- esiste una corrispondenza biunivoca fra coppia (stima, covarianza) e coppia di informazione. In particolare, data la coppia di informazione $\left(\hat{q}_X, \widehat{\Omega}_X\right)$ è possibile determinare univocamente la coppia (stima, covarianza) mediante le relazioni

$$\widehat{\Sigma}_X = \widehat{\Omega}_X^{-1}, \qquad \hat{x} = \widehat{\Omega}_X^{-1}\hat{q}_X.$$

Pertanto (1.42) può essere riscritta in *forma di informazione*

$$\left\{ \begin{array}{rcccl} \widehat{\Omega}_X & = & \Omega_X & + & C^T\Sigma_V^{-1}C \\ \hat{q}_X & = & q_X & + & C^T\Sigma_V^{-1}y \end{array} \right. \tag{1.43}$$

o, più compattamente, come

$$\left[\hat{q}_X, \widehat{\Omega}_X\right] = [q_X, \Omega_X] + C^T\Sigma_V^{-1}[y, C] \tag{1.44}$$

che esprime il fatto che l'informazione a-posteriori $\left[\hat{q}_X, \widehat{\Omega}_X\right]$ risulta dalla somma dell'informazione a-priori $[q_X, \Omega_X]$ e del contributo innovativo $[\delta q_X, \delta\Omega_X] \triangleq C^T\Sigma_V^{-1}[y, C]$ dovuto all'osservazione y.

Riassumendo, il problema di stima lineare può essere risolto analiticamente con i due algoritmi, di covarianza e di informazione, di seguito esposti.

Algoritmo di covarianza

$$
\begin{aligned}
\text{Dati} \quad & \overline{x} \text{ e } \Sigma_X : \\
\Sigma_Y &= \Sigma_V + C\Sigma_X C^T \\
\Sigma_{XY} &= \Sigma_X C^T \\
L &= \Sigma_{XY}\Sigma_Y^{-1} \\
e &= y - C\overline{x} \\
\hat{x} &= \overline{x} + Le \\
\widehat{\Sigma}_X &= (I - LC)\,\Sigma_X\,(I - LC)^T + L\Sigma_V L^T \\
&= \Sigma_X - L\Sigma_Y L^T
\end{aligned}
$$

Si noti che nel precedente algoritmo di covarianza sono state fornite due espressioni matematicamente equivalenti per il calcolo della covarianza a posteriori $\widehat{\Sigma}_X$. Infatti:

$$
\begin{aligned}
\widehat{\Sigma}_X &= \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T \\
&= \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_Y\Sigma_Y^{-1}\Sigma_{XY}^T \\
&= \Sigma_X - \left(\Sigma_{XY}\Sigma_Y^{-1}\right)\Sigma_Y\left(\Sigma_{XY}\Sigma_Y^{-1}\right)^T \\
&= \Sigma_X - L\Sigma_Y L^T \\
&= E\left[(X - \hat{x})(X - \hat{x})^T\right] \\
&= E\left[(X - \overline{x} - Le)(\cdots)^T\right] \\
&= E\left\{[X - \overline{x} - L(Y - C\overline{x})]\,[\cdots]^T\right\} \\
&= E\left\{[X - \overline{x} - L(CX + V - C\overline{x})]\,[\cdots]^T\right\} \\
&= E\left\{[(I - LC)\,(X - \overline{x}) + LV]\,[\cdots]^T\right\} \\
&= (I - LC)E\left[(X - \overline{x})\,(X - \overline{x})^T\right](I - LC)^T + LE\left[VV^T\right]L^T \\
&= (I - LC)\Sigma_X(I - LC)^T + L\Sigma_V L^T.
\end{aligned}
$$

In particolare, la prima espressione, quella che calcola $\widehat{\Sigma}_X$ come somma di due matrici simmetriche semi-definite positive è da preferirsi per ragioni numeriche alla seconda che, viceversa, effettua una differenza fra matrici simmetriche semi-definite positive.

Algoritmo di informazione

$$\begin{aligned}
\text{Dati} \quad & \overline{x} \text{ e } \Sigma_X : \\
\Omega_X &= \Sigma_X^{-1} \\
q_X &= \Omega_X \overline{x} \\
[\delta q_X, \delta \Omega_X] &= C^T \Sigma_V^{-1} [y, C] \\
\left[\hat{q}_X, \widehat{\Omega}_X\right] &= [q_X + \delta q_X, \Omega_X + \delta \Omega_X] \\
\widehat{\Sigma}_X &= \widehat{\Omega}_X^{-1} \\
\hat{x} &= \widehat{\Sigma}_X \, \hat{q}_X
\end{aligned}$$

A conclusione di questo paragrafo, si vuole illustrare un'interpretazione della stima lineare (1.40) come soluzione del problema di ottimizzazione quadratica

$$\hat{x} = \arg\min_x \left\{ \underbrace{(x - \overline{x})^T \Sigma_X^{-1} (x - \overline{x}) + (y - Cx)^T \Sigma_V^{-1} (y - Cx)}_{J(x)} \right\} \tag{1.45}$$

che, in aggiunta al costo ai minimi quadrati $J_{LS}(x) = (y - Cx)^T \Sigma_V^{-1} (y - Cx)$, introduce un termine aggiuntivo $(x - \overline{x})^T \Sigma_X^{-1} (x - \overline{x})$ che penalizza lo scostamento quadratico di x dalla sua media a priori $\overline{x}$ pesato dalla matrice di informazione Σ_X^{-1}. Infatti, sviluppando il costo $J(x)$ in (1.45) si ottiene

$$J(x) = x^T \left(\Sigma_X^{-1} + C^T \Sigma_V^{-1} C \right) x - 2x^T \left(\Sigma_X^{-1} \overline{x} + C^T \Sigma_V^{-1} y \right) + \overline{x}^T \Sigma_X^{-1} \overline{x} + y^T \Sigma_V^{-1} y \tag{1.46}$$

da cui, ricordando che $\frac{\partial}{\partial x} \left(x^T M x \right) = 2Mx$ e $\frac{\partial}{\partial x} \left(x^T z \right) = z$, la soluzione del problema di ottimizzazione (1.46) deve soddisfare

$$\frac{\partial J}{\partial x}(\hat{x}) = 2 \left[\left(\Sigma_X^{-1} + C^T \Sigma_V^{-1} C \right) \hat{x} - \left(\Sigma_X^{-1} \overline{x} + C^T \Sigma_V^{-1} y \right) \right] = 0$$

da cui si ricava immediatamente la stima BLUE (MMSE nel caso Gaussiano)

$$\hat{x} = \left(\Sigma_X^{-1} + C^T \Sigma_V^{-1} C \right)^{-1} \left(\Sigma_X^{-1} \overline{x} + C^T \Sigma_V^{-1} y \right). \tag{1.47}$$

Stima sequenziale nel caso di errori di misura indipendenti

In questa sezione, si vuole affrontare la situazione in cui le osservazioni $y_i \in \mathbb{R}^{p_i}$, $i \geq 1$, si rendono disponibili allo stimatore ad istanti successivi t_i cronologicamente ordinati, i.e. $t_1 < t_2 < \cdots$, e si desidera, in ogni intervallo $[t_i, t_{i+1})$, elaborare la stima $\hat{x}_i$ di X basata sulla osservazioni $y_{1:i} \stackrel{\triangle}{=} \{y_1, y_2, \ldots, y_i\}$ nonché la corrispondente covarianza

$\Sigma_i \triangleq E\left[(X-\hat{x}_i)(X-\hat{x}_i)^T\right]$. In particolare, assumendo che l'acquisizione di nuove osservazioni y_i possa ripetersi indefinitamente nel tempo ($i \to \infty$) o comunque per un numero molto elevato di volte, si vorrebbe effettuare la suddetta elaborazione in modo *sequenziale* o *ricorsivo* nel senso che il calcolo di $(\hat{x}_i, \Sigma_i)$ a partire da $(\hat{x}_{i-1}, \Sigma_{i-1})$ coinvolga un onere computazionale ed occupazione di memoria entrambi indipendenti da i, il numero totale di misure acquisite. Se così non fosse infatti, ovvero l'aggiornamento $(\hat{x}_{i-1}, \Sigma_{i-1}) \to (\hat{x}_i, \Sigma_i)$ richiedesse complessità di calcolo e/o di memoria crescente con i, si arriverebbe ad un certo valore sufficientemente elevato di i per cui non è più possibile completare l'aggiornamento nell'intervallo di tempo disponibile $[t_i, t_{i+1})$.

Sia

$$\begin{cases} Y_i &= C_i X + V_i \qquad i = 1, 2, \dots \\ X &\sim (\overline{x}, \Sigma_X) \\ V_i &\sim (0, R_i) \end{cases}$$

la i-esima osservazione acquisita all'istante t_i dove: $Y_i, V_i \in \mathbb{R}^{p_i}$; $C_i \in \mathbb{R}^{p_i \times n}$; $X \in \mathbb{R}^n$; $R_i \in \mathbb{R}^{p_i \times p_i}$ è una matrice simmetrica definita-positiva, i.e. $R_i = R_i^T > 0$. Nel seguito si mostrerà come sia possibile procedere in modo ricorsivo se gli errori di misura relativi ad osservazioni diverse risultano incorrelati, i.e.

$$V_i \perp V_j = 0, \qquad \forall i \neq j. \tag{1.48}$$

Si fa notare come l'ipotesi (1.48) faccia riferimento ad errori di misura compiuti ad istanti $t_i \neq t_j$ diversi e quindi risulta ragionevole anche nel caso in cui le misure vengono effettuate dallo stesso sensore (ed a maggior ragione se effettuate da sensori diversi).

Si ponga

$$\hat{x}_0 = \overline{x}, \quad \Sigma_0 = \Sigma_X.$$

Questa notazione è peraltro consistente con il fatto che, così come $(\hat{x}_i, \Sigma_i)$ rappresenta (per $i > 0$) la coppia stima-covarianza basata sulle prime i osservazioni, $(\hat{x}_0, \Sigma_0) = (\overline{x}, \Sigma_0)$ è la coppia stima-covarianza a-priori, i.e. basata su nessuna ($i = 0$) osservazione.

Il generico problema di stima i-esimo di X sulla base di $Y_{1:i}$ è descritto in modo compatto dal seguente modello di osservazione

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_i \end{bmatrix} X + \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_i \end{bmatrix}$$

$$\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_i \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} R_1 & 0 & \cdots & 0 \\ 0 & R_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & R_i \end{bmatrix} \right).$$

Utilizzando le formule della stima lineare, si ha

$$\begin{aligned}
\Sigma_i^{-1} &= \Sigma_X^{-1} + \left[C_1^T, C_2^T, \ldots, C_i^T\right] \begin{bmatrix} R_1^{-1} & 0 & \cdots & 0 \\ 0 & R_2^{-1} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & R_i^{-1} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_i \end{bmatrix} \\
&= \Sigma_0^{-1} + \sum_{k=1}^{i} C_k^T R_k^{-1} C_k \\
&= \underbrace{\Sigma_0^{-1} + \sum_{k=1}^{i-1} C_k^T R_k^{-1} C_k}_{\Sigma_{i-1}^{-1}} + C_i^T R_i^{-1} C_i \\
&= \Sigma_{i-1}^{-1} + C_i^T R_i^{-1} C_i
\end{aligned} \tag{1.49}$$

per la matrice di informazione ed analogamente

$$\begin{aligned}
\Sigma_i^{-1}\hat{x}_i &= \Sigma_X^{-1}\overline{x} + \left[C_1^T, C_2^T, \ldots, C_i^T\right] \begin{bmatrix} R_1^{-1} & 0 & \cdots & 0 \\ 0 & R_2^{-1} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & R_i^{-1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} \\
&= \Sigma_0^{-1}\hat{x}_0 + \sum_{k=1}^{i} C_k^T R_k^{-1} y_k \\
&= \underbrace{\Sigma_0^{-1}\hat{x}_0 + \sum_{k=1}^{i-1} C_k^T R_k^{-1} y_k}_{\Sigma_{i-1}^{-1}\hat{x}_{i-1}} + C_i^T R_i^{-1} y_i \\
&= \Sigma_{i-1}^{-1}\hat{x}_{i-1} + C_i^T R_i^{-1} y_i.
\end{aligned} \tag{1.50}$$

Le formule (1.49)-(1.50) forniscono direttamente un algoritmo ricorsivo (sequenziale) di aggiornamento della coppia di informazione $\left(q_i \stackrel{\triangle}{=} \Sigma_i^{-1}\hat{x}_i, \Omega_i \stackrel{\triangle}{=} \Sigma_i^{-1}\right)$. Tale algoritmo, detto *algoritmo sequenziale di informazione* procede come segue.

Algoritmo sequenziale di informazione
Inizializzazione: $\Omega_0 = \Sigma_X^{-1}, \quad q_0 = \Omega_0\overline{x}$

Per $i = 1, 2, \ldots$

$$
\begin{aligned}
q_i &= q_{i-1} + C_i^T R_i^{-1} y_i \\
\Omega_i &= \Omega_{i-1} + C_i^T R_i^{-1} C_i \\
\hat{x}_i &= \Omega_i^{-1} q_i \qquad\qquad \square
\end{aligned}
$$

Si noti che, ad ogni $i \geq 1$, l'algoritmo sopra riportato richiede $O(n^3)$ operazioni ($O(\cdot) =$ dell'ordine di $\cdot$) per la soluzione del sistema di equazioni lineari $\Omega_i \hat{x}_i = q_i$ (l'aggiornamento di q_i e Ω_i richiede un onere inferiore, $O(n^2)$, rispetto ad n) ed una occupazione di memoria $O(n^2)$ per salvare le componenti del vettore q_i e della matrice simmetrica Ω_i. In ogni caso le complessità di calcolo e di memoria, ad ogni i, risultano indipendenti da i, requisito necessario per una elaborazione sequenziale su un orizzonte temporale indefinito.

È possibile anche ottenere, sebbene in maniera meno diretta, un algoritmo sequenziale che aggiorna direttamente la stima $\hat{x}_i$ e la matrice di covarianza Σ_i. A tale proposito, applicando il lemma di inversione di matrice a (1.49), si ottiene la formula di aggiornamento della covarianza:

$$
\Sigma_i = \Sigma_{i-1} - \Sigma_{i-1} C_i^T \left(R_i + C_i \Sigma_{i-1} C_i^T \right)^{-1} C_i \Sigma_{i-1} \tag{1.51}
$$

Per quanto riguarda la stima, da (1.50) si ha:

$$
\begin{aligned}
\hat{x}_i &= \Sigma_i \left(\Sigma_{i-1}^{-1} \hat{x}_{i-1} + C_i^T R_i^{-1} y_i \right) \\
&= \Sigma_i \left(\Sigma_i^{-1} - C_i^T R_i^{-1} C_i \right) \hat{x}_{i-1} + \Sigma_i C_i^T R_i^{-1} y_i \\
&= \hat{x}_{i-1} + \Sigma_i C_i^T R_i^{-1} \left(y_i - C_i \hat{x}_{i-1} \right) \\
&= \hat{x}_{i-1} + \Sigma_{i-1} C_i^T \left(R_i + C_i \Sigma_{i-1} C_i^T \right)^{-1} \left(y_i - C_i \hat{x}_{i-1} \right)
\end{aligned} \tag{1.52}
$$

dove nell'ultima formula si è sfruttato il fatto che

$$
\begin{aligned}
L_i &= \Sigma_i C_i^T R_i^{-1} \\
&= \left[\Sigma_{i-1} - \Sigma_{i-1} C_i^T \left(R_i + C_i \Sigma_{i-1} C_i^T \right)^{-1} C_i \Sigma_{i-1} \right] C_i^T R_i^{-1} \\
&= \Sigma_{i-1} C_i^T \left[I - \left(R_i + C_i \Sigma_{i-1} C_i^T \right)^{-1} C_i \Sigma_{i-1} C_i^T \right] R_i^{-1} \\
&= \Sigma_{i-1} C_i^T \left(R_i + C_i \Sigma_{i-1} C_i^T \right)^{-1} \left(R_i + C_i \Sigma_{i-1} C_i^T - C_i \Sigma_{i-1} C_i^T \right) R_i^{-1} \\
&= \Sigma_{i-1} C_i^T \left(R_i + C_i \Sigma_{i-1} C_i^T \right)^{-1} R_i R_i^{-1} \\
&= \Sigma_{i-1} C_i^T \left(R_i + C_i \Sigma_{i-1} C_i^T \right)^{-1} .
\end{aligned} \tag{1.53}
$$

Pertanto, si possono riassumere i precedenti sviluppi nel seguente *algoritmo sequenziale di covarianza.*

Algoritmo sequenziale di covarianza

Inizializzazione: $\hat{x}_0 = \overline{x}, \quad \Sigma_0 = \Sigma_X$

$$
\begin{aligned}
&Per \quad i = 1, 2, \ldots \\
&S_i = R_i + C_i \Sigma_{i-1} C_i^T \\
&L_i = \Sigma_{i-1} C_i^T S_i^{-1} \\
&e_i = y_i - C_i \hat{x}_{i-1} \\
&\hat{x}_i = \hat{x}_{i-1} + L_i e_i \\
&\Sigma_i = \Sigma_{i-1} - L_i S_i L_i^T = (I - L_i C_i) \Sigma_{i-1} (I - L_i C_i)^T + L_i R_i L_i^T . \qquad \square
\end{aligned}
$$

Si noti che per l'aggiornamento della matrice di covarianza Σ_i sono state riportate nel precedente algoritmo due espressioni matematicamente equivalenti delle quali la seconda, che calcola Σ_i come somma di due matrici simmetriche semi-definite positive, è da preferirsi alla prima per ragioni puramente numeriche. È immediato constatare che, ad ogni i, l'algoritmo sequenziale di covarianza richiede $O(n^2)$ operazioni ed $O(n^2)$ occupazione di memoria.

Stima nel caso di modelli di osservazione non lineare

In questa sezione si vuole porre attenzione al problema della stima parametrica nel caso di generico modello di osservazione non lineare

$$
\begin{cases}
Y = h(X, V) \\
\begin{bmatrix} X \\ V \end{bmatrix} \sim \left(\begin{bmatrix} \overline{x} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_X & 0 \\ 0 & \Sigma_V \end{bmatrix} \right).
\end{cases}
\tag{1.54}
$$

Diversamente dal caso di modello di osservazione lineare precedentemente esaminato non sarà possibile risolvere analiticamente, i.e. in forma chiusa, né il problema di stima BLUE

$$
\begin{aligned}
\hat{x}_{BLUE}(y) &= \overline{x} + \Sigma_{XY} \Sigma_Y^{-1} (y - \overline{y}) \\
\widehat{\Sigma}_{X,BLUE} &= \Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{XY}^T
\end{aligned}
\tag{1.55}
$$

né il problema di stima MMSE

$$\begin{aligned} \hat{x}_{MMSE}(y) &= E[X|Y=y] = \frac{\int x f_{Y|X}(y|x) f_X(x) dx}{\int f_{Y|X}(y|x) f_X(x) dx} \\ \widehat{\Sigma}_{X,MMSE} &= E_{X,Y}\left[\left(X - \widehat{X}(Y)\right)\left(X - \widehat{X}(Y)\right)^T\right] \\ &= \int\int \left(x - \hat{x}_{MMSE}(y)\right)\left(x - \hat{x}_{MMSE}(y)\right)^T f_{Y|X}(y|x) f_X(x)\, dx\, dy. \end{aligned} \tag{1.56}$$

Infatti, per il modello (1.54) non è possibile:

- determinare in modo esatto, a partire da $(\overline{x}, \Sigma_X, \Sigma_V)$, le statistiche del primo e secondo ordine $(\overline{y} = E[Y], \Sigma_Y = var(Y), \Sigma_{XY} = cov(X,Y))$ richieste in (1.55);
- risolvere in forma chiusa gli integrali richiesti in (1.56).

Inoltre, per tale modello, gli stimatori BLUE ed MMSE risultano diversi anche nel caso particolare in cui le variabili X e V siano distribuite in modo Gaussiano. Nel seguito, si mostrerà come effettuare numericamente, in modo approssimato, sia la stima BLUE (1.55) che la stima MMSE (1.56) del parametro X sulla base dell'osservazione Y per il generico modello di osservazione non lineare (1.54) ricorrendo rispettivamente

- alla *trasformata unscented*, nella stima BLUE, per il calcolo approssimato dei momenti di $Y = h(X,V)$ nonché momenti incrociati di X e Y, noti i momenti di X e V;
- al metodo *Monte Carlo*, nella stima MMSE, per una efficiente determinazione numerica degli integrali in (1.56).

Prima di considerare questi due metodi, verrà illustrato un terzo metodo basato sulla linearizzazione del modello di osservazione.

Metodo della linearizzazione

Visto che il problema di stima (BLUE/MMSE nel caso Gaussiano) ammette soluzione esatta, facilmente ottenibile per via algebrica, per un modello di osservazione lineare, un'idea naturale per la soluzione di problemi di stima non lineari è quella di linearizzare il modello di osservazione nell'intorno della stima a-priori disponibile e di applicare il procedimento di stima lineare al modello linearizzato. Poiché la stima a-priori (media) $\overline{x}$ potrebbe essere, a causa di scarsa informazione a-priori, molto lontana dal valore vero incognito di X e, di conseguenza, il modello di osservazione linearizzato una *cattiva*

approssimazione del modello di osservazione non lineare originario, potrebbe essere opportuno procedere iterativamente secondo l'algoritmo di seguito riportato.

Algoritmo iterativo di linearizzazione - Modello di osservazione non-lineare

Scegli numero max. di iterazioni $\overline{k}$ e soglia per il criterio di arresto $\varepsilon > 0$

Inizializzazione: $\hat{x}_0 = \overline{x},\ \widehat{\Sigma}_0 = \Sigma_X,\ k = 0$

Ripeti:

$k = k + 1;$

$C_k = \dfrac{\partial h}{\partial x}(\hat{x}_{k-1}, 0)\,; E_k = \dfrac{\partial h}{\partial v}(\hat{x}_{k-1}, 0)\,;$

$S_k = C_k \widehat{\Sigma}_{k-1} C_k^T + E_k \Sigma_V E_k^T;$

$L_k = \widehat{\Sigma}_{k-1} C_k^T S_k^{-1};$

$e_k = y_k - h(\hat{x}_{k-1}, 0)\,;$

$\hat{x}_k = \hat{x}_{k-1} + L_k e_k;$

$\widehat{\Sigma}_k = \widehat{\Sigma}_{k-1} - L_k S_k L_k^T = (I - L_k C_k)\,\widehat{\Sigma}_{k-1}(I - L_k C_k)^T + L_k \Sigma_V L_k^T$

fino a che $k = \overline{k}$ **o** $(\hat{x}_k - \hat{x}_{k-1})^T \widehat{\Sigma}_k^{-1} (\hat{x}_k - \hat{x}_{k-1}) \leq \varepsilon.$

Stima BLUE mediante metodo della *trasformata unscented*

Date la variabile aleatoria $X \sim (\overline{x}, \Sigma_X)$ e la funzione $g : X \to Y = g(X)$, ci si pone il problema di determinare i momenti $E[Y], var(Y), cov(X, Y)$. Se la funzione $g(\cdot)$ è affine, i.e. della forma $g(X) = CX + b$, la determinazione di tali momenti è immediata, tramite le formule esatte

$$\begin{array}{rcccl} \overline{y} & = & E[Y] & = & C\overline{x} + b \\ \Sigma_Y & = & var(Y) & = & C\Sigma_X C^T \\ \Sigma_{XY} & = & cov(X, Y) & = & \Sigma_X C^T. \end{array}$$

Se, viceversa, la funzione $g(\cdot)$ è non lineare si può soltanto aspirare ad approssimazioni accurate $\overline{y}, \Sigma_Y, \Sigma_{XY}$ di $E[Y], var(Y), cov(X, Y)$.

Uno strumento oramai consolidato a tale scopo è la *trasformata unscented* (UT = *Unscented Transformation*) ideata da Jeffrey Uhlmann nella sua tesi di dottorato del 1995. Per il momento, si considera UT come una procedura di calcolo indicata con la notazione simil-Matlab

$$[\overline{y}, \Sigma_Y, \Sigma_{XY}] \ = \ UT(\overline{x}, \Sigma_X, g(\cdot))$$

che opera sugli argomenti di ingresso $\overline{x}, \Sigma_X, g(\cdot)$ (momenti di X e funzione che lega X ad Y) e restituisce in uscita approssimazioni $\overline{y}, \Sigma_Y, \Sigma_{XY}$ dei momenti desiderati, ignorando

la struttura interna e le proprietà di approssimazione di tale procedura che saranno esaminate successivamente. Per risolvere, in modo approssimato, il problema di stima BLUE per il modello di osservazione (1.54) si può dunque applicare UT alla trasformazione non lineare

$$Y \;=\; g\left(\left[X^T, V^T\right]^T\right) \;\stackrel{\triangle}{=}\; h(X,V)$$

operante sulla variabile aleatoria congiunta $X' = [X^T, V^T]^T$. Ne risulta il seguente algoritmo di stima BLUE utilizzabile per un generico modello di osservazione non lineare.

Algoritmo di stima BLUE mediante UT

$[\overline{y}, \Sigma_Y, \Sigma_{X'Y}] = UT\left(\left[\begin{array}{c} \overline{x} \\ 0 \end{array}\right], \left[\begin{array}{cc} \Sigma_X & 0 \\ 0 & \Sigma_V \end{array}\right], h(\cdot,\cdot)\right)$

estrai Σ_{XY} da $\Sigma_{X'Y} = \left[\begin{array}{c} \Sigma_{XY} \\ \Sigma_{VY} \end{array}\right]$

$L = \Sigma_{XY}\Sigma_Y^{-1}$

$e = y - \overline{y}$

$\hat{x}_{BLUE}(y) = \overline{x} + Le$

$\widehat{\Sigma}_X = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T = \Sigma_X - L\Sigma_Y L^T.$

Si noti che, nel caso particolare di modello di osservazione additivo $Y = h(X) + V$ il precedente algoritmo si semplifica nel seguente modo, applicando UT alla trasformazione $Z = h(X)$ ed osservando che, in virtù della incorrelazione fra X e V, $E[Y] = E[Z], var(Y) = var(Z) + var(V), cov(X,Y) = cov(X,Z)$.

Algoritmo di stima BLUE mediante UT- Modello di osservazione additivo

$[\overline{y}, \Sigma_Z, \Sigma_{XY}] = UT(\overline{x}, \Sigma_X, h(\cdot))$

$\Sigma_Y = \Sigma_Z + \Sigma_V$

$L = \Sigma_{XY}\Sigma_Y^{-1}$

$e = y - \overline{y}$

$\hat{x}_{BLUE}(y) = \overline{x} + Le$

$\widehat{\Sigma}_X = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T = \Sigma_X - L\Sigma_Y L^T.$

A questo punto non rimane che esaminare in dettaglio l'algoritmo UT. L'idea alla base di UT è quella di generare, per la variabile aleatoria $X \sim (\overline{x}, \Sigma_X)$, un insieme finito di campioni x_i, i cosiddetti *punti-sigma*, con relativi pesi ω_i in modo tale che la media e la varianza campionaria su tale insieme coincidano esattamente con $\overline{x}$ e, rispettivamente,

Σ_X, i.e.,

$$\sum_i \omega_i \, x_i \;\; = \;\; \overline{x} \tag{1.57}$$

$$\sum_i \omega_i \, (x_i - \overline{x}) \, (x_i - \overline{x})^T \;\; = \;\; \Sigma_X. \tag{1.58}$$

A questo proposito, si dimostra che sono sufficienti $n{+}1$ campioni dove n è la dimensione di X anche se, per ragioni di simmetria, si preferisce utilizzarne $2n{+}1$ indicati con $x_i \in \mathbb{R}^n$ per $i = -n, \ldots, -1, 0, 1, \ldots, n$ con relativi pesi (non necessariamente non-negativi) $\omega_i \in \mathbb{R}$. Prima di sviluppare la condizione di corrispondenza delle covarianza (1.58), conviene fattorizzare la matrice di covarianza Σ_X nel modo seguente

$$\Sigma_X \;\; = \;\; \Gamma\Gamma^T \;\; = \;\; \sum_{i=1}^{n} \gamma_i \gamma_i^T \tag{1.59}$$

dove γ_i è la i-esima colonna della matrice $\Gamma \in \mathbb{R}^{n \times n}$. Si fa presente come tale fattorizzazione possa essere ottenuta in vari modi, quali ad esempio:

- la fattorizzazione di Cholesky in cui Γ coincide con una matrice triangolare, inferiore o superiore;
- la decomposizione a valori singolari (autovalori-autovettori) $\Sigma_X = V \Lambda V^T$, con V matrice ortogonale degli autovettori e Λ matrice diagonale degli autovalori, da cui $\Gamma = V \Lambda^{1/2}$ (in questo caso γ_i coincidono con autovettori della matrice di covarianza e sono dunque allineati con gli assi di simmetria degli ellissoidi di confidenza);
- la radice quadrata simmetrica $\Gamma = \Gamma^T = \sqrt{\Sigma_X}$.

Sostituendo (1.59) in (1.58), si ottiene la relazione:

$$\sum_{i=-n}^{n} \omega_i \, (x_i - \overline{x}) \, (x_i - \overline{x})^T \;\; = \;\; \sum_{i=1}^{n} \gamma_i \gamma_i^T. \tag{1.60}$$

Una possibile scelta dei punti-sigma per imporre la condizione di corrispondenza delle covarianze (1.60) è quindi la seguente:

$$x_i - \overline{x} = \begin{cases} a_i \gamma_i, & i > 0 \\ 0, & i = 0 \\ -a_i \gamma_i, & i < 0 \end{cases} \implies x_i = \begin{cases} \overline{x} + a_i \gamma_i, & i > 0 \\ \overline{x}, & i = 0 \\ \overline{x} - a_i \gamma_i, & i < 0 \end{cases} \tag{1.61}$$

Infatti, sostituendo (1.61) in (1.60) ed imponendo pesi simmetrici uguali, i.e. $\omega_i = \omega_{-i}$ per $i = 1, \ldots, n$, si ha

$$\sum_{i=1}^{n} 2\omega_i a_i^2 \, \gamma_i \gamma_i^T \;\; = \;\; \sum_{i=1}^{n} \gamma_i \gamma_i^T \tag{1.62}$$

che risulta soddisfatta se si pone, per ogni i, $2\omega_i a_i^2 = 1$ da cui

$$a_i \;=\; \frac{1}{\sqrt{2\omega_i}} \qquad i = 1, \ldots, n. \tag{1.63}$$

Se si sostituisce la scelta dei punti-sigma (1.61) in (1.57), si deduce che la corrispondenza delle medie comporta la seguente relazione

$$\sum_{i=-n}^{n} \omega_i \;=\; 1 \tag{1.64}$$

Riassumendo, la scelta dei punti-sigma (1.61) soddisfa le condizioni (1.57)-(1.58) se si impongono le relazioni (1.63) e (1.64). Per ridurre il numero di gradi di libertà si scelgono tutti i pesi non centrali uguali, i.e.,

$$\omega_{\pm i} \;=\; \omega, \quad i = 1, \ldots, n \tag{1.65}$$

per cui la condizione (1.64) diventa $\omega_0 + 2n\omega = 1$. In questo modo, l'unico parametro da scegliere è il peso centrale ω_0 da cui si ricavano i pesi non centrali

$$\omega_i = \omega = \frac{1 - \omega_0}{2n} \qquad i = \pm 1, \ldots, \pm n \tag{1.66}$$

e, tramite (1.63), i coefficienti

$$a_i = a = \sqrt{\frac{n}{1 - \omega_0}} \qquad i = \pm 1, \ldots, \pm n. \tag{1.67}$$

Nella letteratura sulla trasformata unscented, il peso centrale ω_0 è solitamente espresso nella seguente forma

$$\omega_0 = \frac{\lambda}{n + \lambda} \tag{1.68}$$

in funzione del parametro λ, per cui da (1.66) gli altri pesi (non centrali) risultano

$$\omega_i = \omega = \frac{1}{2(n + \lambda)} \tag{1.69}$$

e da (1.67) si ha

$$a_i = a = \sqrt{n + \lambda}. \tag{1.70}$$

Sostituendo (1.70) in (1.61), i punti-sigma sono dunque definiti nel seguente modo

$$x_{\pm i} = \begin{cases} \overline{x}, & i = 0 \\ \overline{x} \pm \sqrt{n + \lambda}\, \gamma_i, & i = 1, \ldots, n. \end{cases} \tag{1.71}$$

Il passo successivo di UT è quello di generare i punti-sigma trasformati

$$y_i \;=\; g\left(x_i\right) \qquad i = 0, \pm 1, \ldots, \pm n. \tag{1.72}$$

Infine si ottengono $\overline{y}, \Sigma_Y, \Sigma_{XY}$ come medie campionarie sull'insieme dei punti-sigma tramite:

$$\begin{aligned} \overline{y} &= \sum_{i=-n}^{n} \omega_i y_i \\ \Sigma_Y &= \sum_{i=-n}^{n} \omega_i' \left(y_i - \overline{y}\right)\left(y_i - \overline{y}\right)^T \\ \Sigma_{XY} &= \sum_{i=-n}^{n} \omega_i' \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)^T \end{aligned} \tag{1.73}$$

dove, per motivi pratici, si può utilizzare un diverso peso centrale

$$\begin{cases} \omega_0' &= \omega_0 + 1 - \alpha^2 + \beta \\ \omega_i' &= \omega_i \qquad\qquad\qquad\qquad i = \pm 1, \ldots, \pm n \end{cases} \tag{1.74}$$

per il calcolo delle statistiche del secondo ordine. In particolare, la *trasformata unscented scalata* (SUT=Scaled Unscented Transformation) suggerisce la scelta del parametro di scala

$$\lambda \;=\; \alpha^2 \left(n + \kappa\right) - n \tag{1.75}$$

dove i parametri $\alpha \in (0, 1]$ e $\kappa \geq 0$ consentono di influenzare la dispersione dei punti-sigma intorno al loro centro $\overline{x}$. Viceversa, l'altro parametro β in (1.74) è utilizzato per tener conto di eventuale informazione a-priori sulla distribuzione di probabilità di X; in particolare $\beta = 2$ è il valore ottimo, ovvero il valore che rende minima l'entità dell'errore di approssimazione dei momenti, per la distribuzione Gaussiana.

Riassumendo i precedenti sviluppi, si ha il seguente algoritmo per la trasformata unscented.

Algoritmo UT (Trasformata Unscented)

Ingressi: $\overline{x}, \Sigma_X$, funzione $g(\cdot)$

Si scelgono i parametri $\alpha \in (0,1], \kappa \geq 0, \beta \geq 0$ di UT

% Calcolo dei pesi

$\lambda = \alpha^2 (n + \kappa) - n$

$\omega_0 = \dfrac{\lambda}{n + \lambda}$

$\omega_0' = \omega_0 + 1 - \alpha^2 + \beta$

$\omega_i = \omega_i' = \dfrac{1}{2(n+\lambda)} \qquad i = \pm 1, \ldots, \pm n$

% Calcolo dei punti-sigma

Si fattorizza la matrice Σ_X come $\Sigma_X = \Gamma\Gamma^T$ con $\Gamma = [\gamma_1, \gamma_2, \ldots, \gamma_n]$

$x_0 = \overline{x}$

$x_{\pm i} = \overline{x} \pm \sqrt{n + \lambda}\, \gamma_i \qquad i = 1, \ldots, n$

% Calcolo dei punti-sigma trasformati

$y_i = g(x_i) \qquad i = 0, \pm 1, \ldots, \pm n$

% Calcolo dei momenti

$$\overline{y} = \sum_{i=-n}^{n} \omega_i y_i$$

$$\Sigma_Y = \sum_{i=-n}^{n} \omega_i' (y_i - \overline{y}) (y_i - \overline{y})^T$$

$$\Sigma_{XY} = \sum_{i=-n}^{n} \omega_i' (x_i - \overline{x}) (y_i - \overline{y})^T .$$

Uscite: $\overline{y}, \Sigma_Y, \Sigma_{XY}$

Le proprietà più significative della trasformata unscented, studiate in letteratura da Jeffrey Uhlmann e Simon Julier, sono riassunte di seguito.

1. Si può dimostrare che i momenti approssimati $\overline{y}, \Sigma_Y, \Sigma_{XY}$ forniti da UT sono in accordo con i valori veri $E[Y], var(Y), cov(X,Y)$ fino al secondo ordine dello sviluppo di Taylor e quindi più accurati di quelli ottenuti mediante linearizzazione, che lo sono soltanto fino al primo ordine dello sviluppo di Taylor.

2. I punti-sigma catturano la stessa media $\overline{x}$ e covarianza Σ_X a prescindere dal tipo di fattorizzazione adottata per la matrice di covarianza Σ_X. La soluzione più efficiente dal punto di vista computazionale è costituita dalla fattorizzazione di Cholesky.

3. I momenti $\overline{y}, \Sigma_X, \Sigma_{XY}$ sono calcolati utilizzando operazioni standard di vettori e matrici. Questo implica che l'algoritmo UT è applicabile ad ogni funzione $g(\cdot)$, eventualmente discontinua e non derivabile, in quanto non richiede la determinazione di alcun Jacobiano, a differenza del metodo di linearizzazione.

4. UT dispone di tre parametri (gradi di libertà) α, κ, β che consentono una sintonia fine delle prestazioni, in termini di accuratezza, dell'algoritmo. Una possibile scelta è quella di porre $\alpha = 1$ (trasformazione non scalata) e $\beta = 0$ (stessi pesi per la media e la covarianza) da cui deriva $\lambda = \kappa$. In questo caso, la scelta di $\kappa \geq 0$ influenza i termini di ordine superiore al secondo dell'approssimazione dei momenti. In particolare, se si assume che X abbia distribuzione Gaussiana, una buona euristica è quella di porre $n + \kappa = 3$, i.e. $\kappa = 3 - n$. Si noti, tuttavia, che questa scelta comporta un peso centrale $\omega_0 = \omega_0' = \frac{3-n}{3}$ negativo per $n > 3$. A sua volta, $\omega_0' < 0$ potrebbe causare problemi numerici di perdita di positività della matrice Σ_Y. Per ovviare a questo inconveniente, è sufficiente scegliere $\beta \geq \max\left(0, \frac{n}{3} - 1\right)$ in modo da avere $\omega_0' \geq 0$.

Stima MMSE mediante metodo *Monte Carlo*

I metodi Monte Carlo sono stati sviluppati durante la seconda guerra mondiale (anni 1940-1945) da fisici nucleari (Ulam, Fermi, von Neumann, Metropolis) per risolvere problemi della fisica e della matematica, difficili se non impossibili da risolvere per altra via, mediante il ricorso ripetuto alla generazione di campioni/eventi casuali. Con la crescita esponenziale delle potenze di calcolo degli elaboratori elettronici, tali metodi hanno trovato nel corso degli anni crescente impiego nella soluzione di svariati problemi di ottimizzazione, integrazione numerica, simulazione, stima, apprendimento automatico (*machine learning*), etc.. In principio, i metodi Monte Carlo sono utilizzabili per ogni problema che abbia un'interpretazione probabilistica. Nel seguito, si mostrerà in sequenza l'impiego dell'approccio Monte Carlo (detto anche "a particelle") nell'integrazione numerica, per il calcolo dei momenti di una distribuzione di probabilità e per la stima parametrica MMSE.

Integrazione numerica Monte Carlo

Si consideri il problema di calcolare l'integrale multi-dimensionale

$$\mathcal{I} = \int_S f(x)\, dx \tag{1.76}$$

sull'insieme $S \subseteq \mathbb{R}^n$. L'approccio classico di Riemann consiste nel quadrettare la regione S suddividendo ogni suo lato in un certo numero, diciamo k, di parti uguali riconducendo il problema, in ultima analisi, alla valutazione della funzione $f(\cdot)$ in k^n punti, numero che potrebbe facilmente risultare intrattabile anche per valori di k ed n non troppo elevati, e.g. $k = 100$ ed $n = 4$. Un approccio alternativo è quello di generare casualmente N campioni (punti in cui valutare la funzione) con una certa distribuzione di probabilità. A tale proposito, si introduce una PDF $p(\cdot)$, i.e. $\int p(x)dx = 1$, tale che il suo supporto includa la regione S, i.e. tale che $p(x) > 0$ per ogni $x \in S$ o equivalentemente $S_p \stackrel{\triangle}{=} \{x \in \mathbb{R}^n : p(x) > 0\} \supseteq S$. Con semplici passaggi

$$\mathcal{I} \;=\; \int_S \frac{f(x)}{p(x)} p(x)dx \;=\; \int_{S_p} \underbrace{\frac{f'(x)}{p(x)}}_{g(x)} p(x)dx \;=\; \int_{\mathbb{R}^n} g(x)p(x)dx \;=\; E_{p(\cdot)}\left[g(x)\right] \tag{1.77}$$

dove

$$\begin{aligned} f'(x) \;&=\; \begin{cases} f(x), & x \in S \\ 0, & x \notin S \end{cases} \\ g(x) \;&=\; f'(x)/p(x) \end{aligned} \tag{1.78}$$

essendo $f'(x)$ l'estensione nulla della funzione $f(\cdot)$ a tutto $\mathbb{R}^n$. Si noti che (1.77) riconduce il calcolo dell'integrale $\mathcal{I}$ alla determinazione della media della funzione $g(x) = f'(x)/p(x)$ secondo la distribuzione di probabilità caratterizzata dalla PDF $p(\cdot)$. Pertanto si può approssimare tale media, e quindi l'integrale di interesse, mediante media campionaria (aritmetica) della funzione $g(\cdot)$ su un numero N molto elevato di campioni (particelle) indipendenti x_i generati casualmente con la PDF $p(\cdot)$, i.e.,

$$\begin{cases} x_i \sim p(\cdot) \qquad i = 1, 2, \ldots, N \\ \mathcal{I} \cong \mathcal{I}_N = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} g(x_i) = \dfrac{1}{N}\sum_{i=1}^{N} \dfrac{f'(x_i)}{p(x_i)} = \dfrac{1}{N}\sum_{j:x_j \in S} \dfrac{f(x_j)}{p(x_j)}. \end{cases} \tag{1.79}$$

In merito alla qualità dell'approssimazione Monte Carlo $\mathcal{I}_N$ di $\mathcal{I}$ vale il seguente risultato che si basa sulla *legge dei grandi numeri*.

Teorema (Convergenza del metodo Monte Carlo) - Si considerino l'integrale (1.76) e la sua approssimazione Monte Carlo con N particelle (1.79). Se la varianza della funzione $g(\cdot)$ definita in (1.78) secondo la distribuzione di PDF $p(\cdot)$ risulta finita, i.e.,

$$\sigma_g^2 \stackrel{\triangle}{=} E_{p(\cdot)}\left[\left(g(x) - \mathcal{I}\right)^T \left(g(x) - \mathcal{I}\right)\right] = \int \left(g(x) - \mathcal{I}\right)^T \left(g(x) - \mathcal{I}\right) p(x)dx < \infty \tag{1.80}$$

allora l'errore di approssimazione $\mathcal{I} - \mathcal{I}_N$ è distribuito asintoticamente (per $N \to \infty$) in modo normale (Gaussiano) con media nulla e varianza σ_g^2/N, i.e.,

$$\lim_{N\to\infty} \sqrt{N}\left(\mathcal{I} - \mathcal{I}_N\right) \;=\; \mathcal{N}\left(0, \sigma_g^2\right). \tag{1.81}$$

Il precedente risultato non solo asserisce che la stima Monte Carlo $\mathcal{I}_N$ di $\mathcal{I}$ è asintoticamente non polarizzata nel senso che

$$\lim_{N\to\infty} E\left[\mathcal{I} - \mathcal{I}_N\right] = 0 \tag{1.82}$$

ma anche che la varianza di tale stima tende a zero asintoticamente come $1/N$ o, equivalentemente, che l'errore di stima è asintoticamente dell'ordine di $N^{-1/2}$, a prescindere dalla dimensione dello spazio di integrazione. A questo proposito, il metodo Monte Carlo è più efficiente, per $n \geq 3$, del metodo di Riemann che presenta un tasso di convergenza a zero dell'errore di approssimazione dell'ordine di $N^{-1/n}$, dipendente quindi da n.

L'efficienza del metodo Monte Carlo dipende ovviamente anche dalla scelta della PDF $p(\cdot)$, detta *PDF di campionamento*, con la quale si generano le particelle. Si ricorda che il supporto di tale PDF deve contenere il dominio di integrazione S in modo da poter generare particelle che coprano interamente tale dominio senza escludere nessuna sua parte. D'altro canto è opportuno che la PDF $p(\cdot)$ sia facilmente campionabile, nel senso che sia disponibile un generatore di numeri casuali efficiente per estrarre campioni da tale distribuzione. Un altro aspetto importante è che tutti i campioni x_i estratti con $p(\cdot)$ che non appartengono alla regione S non contribuiscono al calcolo di $\mathcal{I}_N$ in quanto per essi risulta $f'(x_i) = 0$. È opportuno dunque, per questioni di efficienza, che il numero di tali campioni sia ridotto al minimo facendo in modo che il supporto di $p(\cdot)$ contenga S in modo aderente, cioè tale che la regione complementare $S_p \backslash S$ sia più piccola possibile. Infine, si fa notare che l'approssimazione Monte Carlo è proporzionale alla varianza della funzione $g(x) = f'(x)/p(x)$ per cui un ulteriore criterio di scelta della PDF di campionamento $p(\cdot)$ è quello di minimizzare, o comunque cercare di ridurre, tale varianza.

Per esemplificare il procedimento di integrazione Monte Carlo sopra esposto, se ne considera l'applicazione al calcolo del volume di una certa regione S, i.e. dell'integrale

$$\mathcal{I} \;=\; vol(S) \stackrel{\triangle}{=} \int_S dx.$$

A tale scopo, si può ipotizzare di conoscere una regione Ω di forma e volume noto che racchiude S nel modo più aderente possibile ed utilizzare una distribuzione di campionamento uniforme in Ω, i.e.,

$$p(x) \;=\; \begin{cases} \dfrac{1}{vol(\Omega)}, & x \in \Omega \\ 0, & x \notin \Omega. \end{cases}$$

Conseguentemente, l'approssimazione Monte Carlo del volume è data da

$$vol(S) \cong \mathcal{I}_N \;=\; \frac{vol(\Omega)}{N} \sum_{i=1}^{N} f'(x_i) \;=\; \frac{N(S)}{N}\; vol(\Omega) \tag{1.83}$$

dove: le particelle $\{x_i\}_{i=1}^N$ sono generate con distribuzione uniforme in Ω, i.e. $x_i = \mathcal{U}(\Omega)$; $N(S)$ è il numero di tali particelle contenute in $S \subset \Omega$. In virtù del teorema di convergenza del metodo Monte Carlo si ha che $\lim_{N\to\infty} \mathcal{I}_N = vol(S)$ da cui

$$\lim_{N\to\infty} \frac{N(S)}{N} = \frac{vol(S)}{vol(\Omega)} \tag{1.84}$$

che esprime il fatto che la frazione di punti generati contenuta in S tende, per $N \to \infty$, al rapporto fra i volumi di S e di Ω.

Riassumendo, si può calcolare in modo approssimato il volume di S procedendo come segue.

1. Si sceglie una regione Ω, di volume $vol(\Omega)$ noto, che contenga S. Ad esempio, posta l'origine in un punto interno centrale di S, si sceglie Ω come l'ipercubo centrato nell'origine e tangente esternamente ad S, i.e.,

   $$\Omega = \{x \in \mathbb{R}^n : \|x\|_\infty \leq \overline{x}\} \quad \text{con} \quad \overline{x} = \sup_{x\in S} \|x\|_\infty$$

   di volume $vol(\Omega) = (2\overline{x})^n$.

2. Si generano casualmente N campioni con distribuzione uniforme in Ω.

3. Per ogni campione generato si verifica la sua appartenenza ad S e si conta il numero totale $N(S)$ di campioni in S.

4. Si determina l'approssimazione $\mathcal{I}_N$ di $vol(S)$ mediante (1.83).

Ad esempio, per calcolare il volume dell'ipersfera S di raggio unitario in $\mathbb{R}^n$ si potrebbero generare casualmente $N = O(10^4)$ campioni nell'ipercubo $\Omega = \{x : \|x\|_\infty \leq 1\}$ che circoscrive S, di volume $vol(\Omega) = 2^n$, contare il numero di campioni x_i contenuti in S, i.e.,

$$N(S) = \left|\left\{i \in \{1, \ldots, N\} : \|x_i\|_2^2 = x_i^T x_i \leq 1\right\}\right|$$

dove $|\{\cdots\}|$ indica la *cardinalità* (numero di elementi) dell'insieme $\{\cdots\}$, ed infine utilizzare la formula (1.83).

Calcolo dei momenti di una distribuzione di probabilità

Come applicazione dell'integrazione numerica Monte Carlo si vuole considerare il problema della determinazione dei momenti di una certa variabile aleatoria $X \sim p(\cdot)$, i.e. del calcolo di integrali del tipo

$$\mathcal{I} = E[g(X)] = \int g(x)p(x)dx.$$

Seguendo l'impostazione della precedente sezione si potrebbero estrarre N campioni x_i dalla PDF $p(\cdot)$ e con essi valutare in modo approssimato $E[g(X)]$ mediando aritmeticamente $g(x_i)$ per $i = 1, \ldots, N$. Per varie ragioni, tuttavia, potrebbe essere preferibile usare una diversa PDF $q(\cdot)$, detta *PDF di proposta*, al posto di $p(\cdot)$ per la generazione dei campioni di X. A tale proposito, si ricorre al seguente artificio

$$\mathcal{I} = E_{p(\cdot)}[g(X)] = \int g(x) \underbrace{\frac{p(x)}{q(x)}}_{w(x)} q(x)\, dx = E_{q(\cdot)}[w(X)g(X)] \tag{1.85}$$

che essenzialmente riconduce il problema del calcolo della media della funzione $g(\cdot)$ secondo la PDF originaria $p(\cdot)$ al calcolo della media della funzione modificata $w(\cdot)g(\cdot)$ secondo la PDF di proposta $q(\cdot)$. In questo modo, risulta possibile approssimare il momento di interesse $\mathcal{I}$ con campioni estratti da $q(\cdot)$ procedendo come segue:

$$\begin{cases} x_i \sim q(\cdot) & i = 1, 2, \ldots, N \\ \mathcal{I} \cong \mathcal{I}_N = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} w(x_i)\, g(x_i). \end{cases} \tag{1.86}$$

Si noti come la funzione $w(x) \stackrel{\triangle}{=} p(x)/q(x)$, detta *peso di importanza*, serve a compensare il fatto che si campiona la variabile aleatoria X con una PDF di proposta $q(\cdot)$ diversa da quella, $p(\cdot)$, con cui X è effettivamente distribuita. L'unico vincolo cui deve sottostare la scelta della PDF di proposta $q(\cdot)$ è che il suo supporto deve contenere quello di $p(\cdot)$ in modo che la generazione dei campioni con $q(\cdot)$ non escluda nessun valore possibile di X. Deve pertanto valere che $S_p = \{x : p(x) > 0\} \subseteq S_q = \{x : q(x) > 0\}$ ovvero che $p(x) > 0 \Rightarrow q(x) > 0$ o anche che $q(x) = 0 \Rightarrow p(x) = 0$.

Stima MMSE

Si vuole, infine, mostrare come applicare il metodo Monte Carlo alla stima MMSE, ovvero al calcolo di momenti condizionati del tipo

$$\begin{aligned} \mathcal{I} &= E[g(X)|Y = y] = \int g(x) f_{X|Y}(x|y) dx \\ &= \frac{\displaystyle\int g(x) f_{Y|X}(y|x) f_X(x) dx}{\displaystyle\int f_{Y|X}(y|x) f_X(x) dx} \\ &= \frac{\displaystyle\int g(x) \frac{f_{Y|X}(y|x) f_X(x)}{q(x)} q(x) dx}{\displaystyle\int \frac{f_{Y|X}(y|x) f_X(x)}{q(x)} q(x) dx} \end{aligned} \tag{1.87}$$

dove è stata introdotta una PDF di proposta $q(\cdot)$ che deve essere tale che il suo supporto contenga quello della PDF a-priori $f_X(\cdot)$, i.e. $f_X(x) > 0 \Rightarrow q(x) > 0$. Sfruttando il metodo Monte Carlo per l'approssimazione numerica dei due integrali, a numeratore e a denominatore, in (1.87) e definendo la funzione *peso di importanza*

$$w(x) \;=\; \frac{f_{Y|X}(y|x) f_X(x)}{q(x)} \tag{1.88}$$

si ottiene la seguente approssimazione del momento condizionato di interesse:

$$\begin{cases} x_i \sim q(\cdot) \qquad\qquad i = 1, 2, \ldots, N \\ \\ \mathcal{I} \cong \mathcal{I}_N = \dfrac{\displaystyle\sum_{i=1}^{N} w(x_i)\, g(x_i)}{\displaystyle\sum_{j=1}^{n} w(x_j)} \;=\; \displaystyle\sum_{i=1}^{N} \overline{w}_i\, g(x_i) \end{cases} \tag{1.89}$$

dove

$$\overline{w}_i \;=\; \overline{w}(x_i) \;=\; \frac{w(x_i)}{\displaystyle\sum_{j=1}^{N} w(x_i)} \tag{1.90}$$

sono i pesi di importanza normalizzati in modo tale che

$$\sum_{i=1}^{N} \overline{w}_i \;=\; 1. \tag{1.91}$$

In particolare, se si desidera approssimare la stima MMSE $\hat{x}_{MMSE} = E[X|Y = y]$ si pone $g(x) = x$ e si ha:

$$\hat{x}_{MMSE}(y) = E[X|Y = y] \cong \hat{x} = \sum_{i=1}^{N} \overline{w}_i\, x_i \tag{1.92}$$

ed in modo analogo per il corrispondente MMSE si pone $g(x) = (x - \hat{x}_{MMSE})(x - \hat{x}_{MMSE})^T$ ottenendo

$$\widehat{\Sigma}_{X,MMSE}(y) = E\left[(X - \hat{x}_{MMSE})(\cdots)^T | Y = y\right] \cong \widehat{\Sigma}_X = \sum_{i=1}^{N} \overline{w}_i\, (x_i - \hat{x})(x_i - \hat{x})^T.$$

Riassumendo i precedenti sviluppi, si deduce il seguente algoritmo per la stima MMSE mediante metodo Monte Carlo.

Algoritmo di stima parametrica MMSE mediante metodo Monte Carlo

Si scelgono il numero di particelle N e la densità di proposta $q(\cdot)$ in modo tale che $S_q \supseteq S_{f_X}$

% Generazione dei campioni

$x_i \sim q(\cdot) \qquad i = 1, \ldots, N$

% Calcolo dei pesi di importanza

$$w_i \;=\; \frac{f_{Y|X}(y|x_i) f_X(x_i)}{q(x_i)} \qquad i = 1, \ldots, N$$

% Normalizzazione dei pesi

$$\overline{w}_i = \frac{w_i}{\sum_{j=1}^{N} w_j}$$

% Calcolo di stima MMSE e relativo MSE

$$\hat{x} = \sum_{i=1}^{N} \overline{w}_i x_i$$

$$\widehat{\Sigma}_X = \sum_{i=1}^{N} \overline{w}_i \left(x_i - \hat{x}\right)\left(x_i - \hat{x}\right)^T .$$

Si noti che la PDF condizionata $f_{X|Y}(\cdot|\cdot)$ utilizzata nel calcolo dei pesi di importanza non è altro che la funzione di verosimiglianza relativa al modello di osservazione $Y = h(X, V)$. Si ricorda che, se la funzione di misura $h(x, v)$ è invertibile rispetto al secondo argomento v, si può determinare la verosimiglianza mediante

$$f_{Y|X}(y|x) \;=\; \frac{f_V\left(h^{-1}(x, y)\right)}{\left| det\, \dfrac{\partial h}{\partial v}\left(x, h^{-1}(x, y)\right)\right|}$$

o più semplicemente, nel caso di modello di osservazione additivo,

$$f_{Y|X}(y|x) \;=\; f_V\left(y - h(x)\right). \tag{1.93}$$

Se inoltre l'errore di misura V è Gaussiano, (1.93) diventa

$$f_{Y|X}(y|x) = f_V(y - h(x)) = \frac{1}{\sqrt{(2\pi)^n det\, \Sigma_V}}\, exp\left[-(y - h(x))^T \Sigma_V^{-1} (y - h(x))\right]. \tag{1.94}$$

Per quanto riguarda la scelta della PDF di proposta, questa risulta ovviamente fondamentale per l'accuratezza della stima in funzione del numero N di particelle. La scelta più naturale e più semplice è quella di porre

$$q(x) \;=\; f_X(x)$$

da cui la funzione peso di importanza (1.88) si riduce alla funzione di verosimiglianza, i.e., $w(x) = f_{Y|X}(y|x)$. Questa scelta, tuttavia, può indurre il ben noto problema di *degenerazione dei pesi* in modo tanto più accentuato quanto più *appuntita* è la funzione di verosimiglianza. Per illustrare il problema si consideri il caso di sensori molto accurati con errore di misura V distribuito in modo Gaussiano e di varianza molto piccola, i.e. $\Sigma_V \cong 0$. In questo caso, la funzione di verosimiglianza risulta Gaussiana, i.e. della forma (1.94), con probabilità concentrata in un intorno molto piccolo di $v = y - h(x) \cong 0$. Conseguentemente, la verosimiglianza assume valori molto piccoli quasi dappertutto in $\mathbb{R}^n$ per cui i campioni x_i assumono con elevata probabilità pesi w_i quasi nulli. Quest'ultimo fenomeno va sotto il nome di *degenerazione dei pesi*; in ultima analisi, dopo la normalizzazione dei pesi, potrebbe accadere che:

$$\exists j \in \{1, \ldots, N\} : \overline{w}_j = 1 \text{ e } \overline{w}_i = 0, \forall i \neq j.$$

In ogni caso, l'effetto deleterio della degenerazione dei pesi è che soltanto pochi (al limite uno solo) di essi contribuiscono al calcolo della stima e della relativa covarianza. Si può ovviare a questo inconveniente con opportuni accorgimenti quali, ad esempio, una scelta appropriata della PDF di proposta oppure mediante ricampionamento (*resampling*).

Il ricampionamento consiste di una ri-generazione dei campioni con probabilità discreta definita dai pesi di importanza. Più precisamente, siano $\{(x_i, \overline{w}_i)\}_{i=1}^N$ le particelle prime del ricampionamento e $\{(x'_i, \overline{w}'_i)\}_{i=1}^N$ quelle dopo il ricampionamento, allora l'obiettivo è che

$$\overline{w}'_j = \frac{1}{N} \text{ e } x'_j = x_i \text{ con probabilità } \overline{w}_i \text{ per } j = 1, \ldots, N.$$

In realtà, per evitare l'*impoverimento dei campioni* i.e. la presenza di molte repliche dello stesso campione, si introduce una perturbazione casuale sui campioni rigenerati tramite

$$\begin{cases} x'_j &= x_i + \xi_j \\ \xi_j &= \mathcal{N}(0, \varepsilon I), \varepsilon > 0 \end{cases}$$

in modo da forzare la *diversità* dei campioni x'_j. Riassumendo quanto sopra esposto, si ha il seguente algoritmo di ricampionamento.

Algoritmo di ricampionamento

Ingressi: $\{(x_i, \overline{w}_i)\}_{i=1}^{N}, \varepsilon > 0$

$s_1 = \overline{w}_1$

for $i = 2 : N$

$s_i = s_{i-1} + \overline{w}_i$

end

for $j = 1 : N$

$u = \mathcal{U}([0, 1])$

$\xi_j = \mathcal{N}(0, \varepsilon I)$

if $u \in [s_{i-1}, , s_i]$ then $x'_j = x_i + \xi_j$

$\overline{w}'_j = \dfrac{1}{N}$

end

Uscite: $\left\{\left(x'_j, \overline{w}'_j\right)\right\}_{j=1}^{N}$

Il ricampionamento può essere effettuato ripetutamente, ad esempio suddividendo il vettore delle osservazioni y in k parti i.e. $y = [y_1^T, \ldots, y_k^T]^T$ con errori di misura indipendenti, e ricampionando dopo l'elaborazione di ogni osservazione y_i per $i = 1, \ldots, k-1$. Sia $y_i = h_i(X) + v_i$, allora l'assunzione di indipendenza consente l'elaborazione sequenziale di $y_1, y_2, \ldots, y_k$ in quanto

$$f_{Y|X}(y|x) = f_V(y - h(x)) = \prod_{i=1}^{k} f_{V_i}(y_i - h_i(x)) = \prod_{i=1}^{k} f_{Y_i|X}(y_i|x).$$

Di seguito si riporta un possibile algoritmo di stima parametrica MMSE Monte Carlo con ricampionamento.

Algoritmo di stima MMSE Monte Carlo con ricampionamento

$x'_j \sim f_X(\cdot) \qquad j = 1, \ldots, N$

$w_j = f_{Y_1|X}(y_1|x_j) \qquad j = 1, \ldots, N$

$\overline{w}_j = w_j \left(\sum_{h=1}^{N} w_h \right)^{-1} \qquad j = 1, \ldots, N$

for $i = 2 : k$

Resampling: $\{(x_j, \overline{w}_j)\}_{j=1}^{N} \longrightarrow \left\{ \left(x'_j, \overline{w}'_j\right) \right\}_{j=1}^{N}$

$x_j \leftarrow x'_j \qquad j = 1, \ldots, N$

$w_j = f_{Y_i|X}(y_i|x_j) \qquad j = 1, \ldots, N$

$\overline{w}_j = w_j \left(\sum_{h=1}^{N} w_h \right)^{-1} \qquad j = 1, \ldots, N$

end

$\hat{x} = \sum_{i=1}^{N} \overline{w}_i x_i$

$\widehat{\Sigma}_X = \sum_{i=1}^{N} \overline{w}_i \left(x_i - \hat{x}\right) \left(x_i - \hat{x}\right)^T .$